IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Application of: Turner, Jr. *et al.*

Serial No.: 09/770,643 Group Art Unit: 1647

Filed: 01/26/2001 Examiner: R. Landsman

For: Polynucleotides Encoding Human Neurexin-Like Proteins (As Amended) Attorney Docket No.: LEX-0122-USA

APPEAL BRIEF

Mail Stop Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

TABLE OF CONTENTS

APPEAL BRIEF

Sir:

Appellants hereby submit an original and two copies of this Appeal Brief to the Board of Patent Appeals and Interferences ("the Board") in response to the Final Office Action mailed on October 22, 2002. The Notice of Appeal was timely submitted on February 21, 2003, and was received in the Patent and Trademark Office ("the Office") on February 28, 2003. This Appeal Brief is timely submitted in light of the concurrently filed Petition for an Extension of Time of three months to and including July 28, 2003, and authorization to deduct the fee as required under 37 C.F.R. § 1.17(a)(3) from Appellants' Representatives' deposit account. The Commissioner is also authorized to charge the fee for filing this Appeal Brief ($160.00), as required under 37 C.F.R. § 1.17(c), to Lexicon Genetics Incorporated Deposit Account No. 50-0892.

Appellants believe no fees in addition to the fee for filing the Appeal Brief and the fee for the extension of time are due in connection with this Appeal Brief. However, should any additional fees under 37 C.F.R. §§ 1.16 to 1.21 be required for any reason related to this communication, the Commissioner is authorized to charge any underpayment or credit any overpayment to Lexicon Genetics Incorporated Deposit Account No. 50-0892.

I. REAL PARTY IN INTEREST

The real party in interest is the Assignee, Lexicon Genetics Incorporated, 8800 Technology Forest Place, The Woodlands, Texas, 77381.

II. RELATED APPEALS AND INTERFERENCES

Appellants know of no related appeals or interferences.

III. STATUS OF THE CLAIMS

The present application was filed on January 26, 2001, claiming the benefit of U.S. Provisional

Application Number 60/178,557, which was filed on January 26, 2000, and U.S. Provisional Application Number 60/199,513, which was filed on April 25, 2000, and included original claims 1-5. A Restriction and Election Requirement was set forth during a telephone interview between the Examiner and Appellants' representative David Hibler on February 8, 2002, separating the original claims into three separate and distinct inventions. During this telephone conference, Appellants provisionally elected without traverse the claims of the Group I invention (original claims 1-3) for prosecution on the merits.

A First Official Action on the merits ("the First Action") was issued on April 23, 2002, in which the title and abstract of the application, the Oath/Declaration, and claims 1 and 2 were objected to, claims 1-3 were rejected under 35 U.S.C. § 101 as allegedly lacking a patentable utility, claims 1-3 were rejected under 35 U.S.C. § 112, first paragraph, as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility, claim 1 was rejected under 35 U.S.C. § 112, first paragraph, as allegedly lacking enablement for the full scope of the claimed invention, claim 1 was rejected under 35 U.S.C. § 112, first paragraph, as allegedly not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor, at the time the application was filed, had possession of the claimed invention, claim 2 was rejected under 35 U.S.C. § 112, second paragraph, as allegedly indefinite, and claim 1 was rejected under 35 U.S.C. § 102(b) as allegedly anticipated by Hillier *et al.* (Accession Number AA069426). In a response to the First Official Action submitted to the Office on September 20, 2002 ("response to the First Action"), Appellants submitted a supplemental declaration, amended the title of the application, cancelled claims 4 and 5 without prejudice and without disclaimer as drawn to non-elected inventions, amended claims 1 and 2, added new claims 6-8, and addressed the rejections of claims 1-3.

A Second and Final Official Action ("the Final Action") was mailed on October 22, 2002, indicating that the objections to the title and abstract of the application, the Oath/Declaration, and claims 1 and 2, and the rejections of claim 1 under 35 U.S.C. § 112, first paragraph, as allegedly lacking enablement for the full scope of the claimed invention, claim 1 under 35 U.S.C. § 112, first paragraph, as allegedly not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor, at the time the application was filed, had possession of the claimed invention, claim 2 under 35 U.S.C. § 112, second paragraph, as allegedly indefinite, and claim 1 under 35 U.S.C. § 102(b) as allegedly anticipated by Hillier *et al.* (Accession Number AA069426), had been overcome by the amendments and remarks submitted in the response to the First Action, but maintaining the rejection of claims 1-3 (and newly added claims 6-8) under 35 U.S.C. § 101 as allegedly lacking a patentable utility and under 35 U.S.C. § 112, first paragraph as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility. In a response to the Second and Final Office Action submitted on January 22, 2003 ("response to the Final Action"), Appellants again addressed the rejections of claims 1-3 and 6-8. An Advisory Action ("the Advisory Action") was mailed on February 19, 2003, maintaining the rejection of claims 1-3 and 6-8 under 35 U.S.C. § 101 as allegedly lacking a patentable utility and under 35 U.S.C. § 112, first paragraph as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility. Therefore, claims 1-3 and 6-8 are the subject of this appeal. A copy of the appealed claims are included below in the Appendix (Section IX).

IV. STATUS OF THE AMENDMENTS

As no amendments subsequent to the Final Action have been filed, Appellants believe that no outstanding amendments exist.

V. SUMMARY OF THE INVENTION

The present invention relates to Appellants' discovery and identification of novel human polynucleotide sequences that encode proteins sharing sequence similarity with animal neurexin proteins, particularly contactin associated proteins (see, at least, the specification at page 1, lines 9-12 and page 15, lines 18-21).

The presently claimed polynucleotide sequences were compiled from clustered human gene trapped sequences, ESTs, and cDNAs from human brain, fetal brain, cerebellum and hypothalamus cDNA libraries (specification at page 15, lines 15-17).

The specification details a number of uses for the presently claimed polynucleotide sequences, including assessing gene expression patterns, particularly in diagnostic assays such as forensic analysis (see, for example, the specification at page 10, lines 15-19), in expression profiling using a high throughput "chip" format (specification at page 5, lines 12-14), in determining the genomic structure, for example through the identification of coding sequence, and mapping the sequences to a specific region of a human chromosome (specification at page 10, line 18).

VI. ISSUES ON APPEAL

1. Do claims 1-3 and 6-8 lack a patentable utility?
2. Are claims 1-3 and 6-8 unusable by a skilled artisan due to a lack of patentable utility?

VII. GROUPING OF THE CLAIMS

For the purposes of the outstanding rejections under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph concerning utility, the claims will stand or fall together.

VIII. ARGUMENT

A. Do Claims 1-3 and 6-8 Lack a Patentable Utility?

The Final Action next rejects claims 1-3 and 6-8 under 35 U.S.C. § 101, as allegedly lacking a patentable utility due to not being supported by either a specific and substantial utility or a well-established utility.

The Final Action admits that contactin associated proteins (casprs) have a "specific utility" (the Final Action at page 3), but that "Applicants have not provided any specific information regarding the specific utility of the proteins of the present invention which distinguishes them from other members of the neurexin superfamily" (the Final Action at page 3). Appellants respectfully point out that the presently claimed sequence is clearly referred to in the specification as originally filed as a contactin associated protein (see, at least, the specification at page 1, lines 9-12, and page 15, lines 18-21). Furthermore, as set forth in the Response to the First Action and the Response to the Final Action, two sequences sharing nearly 100% percent identity at the protein level over the entire length of the claimed sequence are present in the leading scientific repository for biological sequence data (GenBank), and have been annotated by third party scientists *wholly unaffiliated with Appellants* as "Homo sapiens caspr5 protein" (GenBank accession numbers NM_130773 (alignment and GenBank report provided in **Exhibit A**) and AB077881 (alignment and GenBank report provided in **Exhibit B**)). The legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be credible or believable. Given these GenBank annotations, there can be no question that those skilled in the art would clearly believe that Appellants' sequence is a caspr protein.

Furthermore, it is well-known in the art that caspr proteins are **distinct members** of the neurexin superfamily (see Poliak *et al.*, Neuron *24*:1037-1047, 1999, and Spiegel *et al.*, Mol. Cell. Neurosci. *20*:283-297, 2002; abstracts provided in **Exhibit C**). As a matter of law, it is well settled that a patent need not disclose what is well known in the art. *In re Wands*, 8 USPQ 2d 1400 (Fed. Cir. 1988). Additionally, as described in the Response to the Final Action, the previously described caspr proteins (caspr 2, 3 and 4) share between 42% and 63% homology with each other (caspr 2 (GenBank accession number AAF25199) vs. caspr 3 (GenBank accession number NP_387504), 42% identity (alignment provided in **Exhibit D**), caspr 2 vs. caspr 4 (GenBank accession number NP_207837), 48% identity (**Exhibit E**), and caspr 4 vs. caspr 3, 63% identity (alignment provided in **Exhibit F**)), but only 23% to 26% homology to neurexins (neurexin 1, 2 and 3); neurexin 1 (GenBank accession number NP_004792) vs. caspr 2, 24% identity (alignment provided in **Exhibit G**), neurexin 1 vs. caspr 3, 23% identity (Exhibit H), neurexin 1 vs. caspr 4, 25% identity (alignment provided in **Exhibit I**), neurexin 2 (GenBank accession number NP_620060) vs. caspr 2, 24% identity (alignment provided in **Exhibit J**), neurexin 2 vs. caspr 3, 25% identity (alignment provided in **Exhibit K**), neurexin 2 vs. caspr 4, 26% identity (alignment provided in **Exhibit L**), neurexin 3 (GenBank accession number CAC87720) vs. caspr 2, 26% identity (alignment provided in **Exhibit M**), neurexin 3 vs, caspr 3, 25% identity (alignment provided in **Exhibit N**), neurexin 3 vs. caspr 4, 25% identity (alignment provided in **Exhibit O**). That Appellants claimed sequence is a caspr is further confirmed by the fact that Appellants sequence shares between 48% and 59% homology to the other caspr proteins (vs. caspr 2, 51% identity (alignment provided in **Exhibit P**), vs. caspr 3, 48% identity (alignment provided in **Exhibit Q**), and vs, caspr 4, 59% identity (alignment provided in **Exhibit R**)), but only 24% to 26% homology to the neurexin proteins (vs. neurexin 1, 25% identity (alignment provided in **Exhibit S**), vs. neurexin 2, 24% identity (alignment provided in Exhibit T), and vs. neurexin 3, 26% identity (alignment provided in **Exhibit U**)), perfectly in line with the previously established figures. Given these data, there can be no question that those skilled in the art would clearly believe that Appellants' sequence is a caspr protein, as opposed to "other members of the neurexin superfamily". As the Examiner admits that casprs have a specific utility, due to their association with "myelinated axons and potassium channels" (the Final Action at page 3), the claimed sequence clearly meets the requirements of 35 U.S.C. § 101.

Nevertheless, the Advisory Action continues to question Appellants asserted utility, stating that "Applicants have only alleged that the protein of the present invention is a caspr based on homology to neurexins and casprs" (Advisory Action at page 2). Appellants respectfully point out that this is **all that is required** for the claimed sequence to meet the requirements of 35 U.S.C. § 101. The present situation **directly** tracks Example 10 of the Revised Interim Utility Guidelines Training Materials (pages 53-55; **Exhibit V**), which **clearly** establishes that a rejection under 35 U.S.C. § 101 as allegedly lacking a patentable utility and under 35 U.S.C. § 112, first paragraph as allegedly unusable by the skilled artisan due to the alleged lack of patentable utility (see Section VIII(B), below), is not proper when a full length sequence (such as the presently claimed sequence), and has a similarity score greater than 95% to a protein having a known function (such as the nearly 100% identity between the presently claimed sequence and the caspr 5 sequences, as discussed above). The Advisory Action concludes that "Applicants have not provided any definitive evidence or statement concluding that the protein of the invention is, in fact, a caspr protein as opposed to another member of the neurexin family of proteins" (Advisory Action at page 2). As discussed at length in the previous paragraph, this assertion is simply not true. Appellants have repeatedly provided evidence and stated for the record that the presently claimed sequence encodes a caspr protein. The Examiner seems to be focusing on Appellants statements in the specification that the presently claimed sequence "share sequence similarity with animal neurexin proteins and contactin associated proteins" (specification at page 1, lines 11-12) and share "similarity with a variety of proteins, including, but not limited to, neurexins (including secreted types) and contactin associated proteins" (specification at page 15, lines 19-21) as an admission that "Applicants have not provided any definitive evidence or statement concluding that the protein of the invention is, in fact, a caspr protein as opposed to another member of the neurexin family of proteins". However, the statements in the specification as originally filed are completely correct, in that as contactin associated proteins are well known to be members of the neurexin superfamily, the presently claimed sequence, as encoding a caspr protein, does in fact share similarity with neurexin proteins. Furthermore, by **specifically** singling out that the presently claimed sequence shares similarity with contactin associated proteins, Appellants leave no doubt that the presently claimed sequence specifically is a caspr protein, as opposed to any of the other members of the neurexin superfamily. Thus, the Examiner's arguments in **no way** support the allegation that the presently claimed sequence lacks a patentable utility.

Although not reiterated in the Final Action or in the Advisory Action, in the First Action the Examiner questioned Appellants' assertion that the presently claimed sequence encodes a caspr protein, citing a number of scientific articles to support this position. Although this issue has been overcome above, in an abundance of caution Appellants wish to take this opportunity to refute the points raised by the Examiner in the First Action. The First Action cited an article by Skolnick *et al.* ("Skolnick"; 2000, Trends in Biotech. 18:34-39) for the proposition that "(k)nowing the protein structure by itself is insufficient to annotate a number of functional classes and is also insufficient for annotating the specific details of protein function" (Skolnick at page 36, emphasis added). However, Skolnick concerns predicting protein function not by overall amino acid homology to other family members, but instead concerns prediction of function based on the presence of certain functional "motifs" present within a given protein sequence. Thus, Skolnick does not apply to the current situation, where overall protein homology is used to assign function to a particular sequence. However, even in the event that Skolnick is applicable, Skolnick itself concludes that "sequence-based approaches to protein-function prediction have proved to be very useful" (Skolnick at page 37), admitting that such methods have correctly assigned function in 50-70% of the cases, thus arguing against the conclusion drawn by the Examiner in the First Action.

The Examiner next cited Bork (Genome Research *10*:398-400, 2000) as supporting the proposition that prediction of protein function from homology information is somewhat unpredictable. However, nowhere in Bork is there a comparison of the prediction accuracy based on the percentage homology between two proteins or two classes of proteins, and thus does not support the alleged lack of utility for the present invention. Additionally, Bork concludes that "there is still no doubt that sequence analysis is extremely powerful" (Bork at page 400), also arguing against the conclusion drawn by the Examiner in the First Action.

The Examiner next cited Doerks *et al.* (Trends in Genetics *14*:248-250, 1998) for the proposition that sequence-to-function methods of assigning protein function are prone to errors. However, Doerks *et al.* states that "utilization of family information and thus a more detailed characterization" should lead to "simplification of update procedures for the entire families if functional information becomes available for at least one member" (Doerks *et al.*, page 248, paragraph bridging columns 1 and 2, emphasis added). Appellants point out that, as detailed above, two sequences sharing nearly 100% percent identity at the protein level with the claimed sequence are present in the leading scientific repository for biological sequence data (GenBank), and have been annotated by third party scientists *wholly unaffiliated with Appellants* as "Homo sapiens caspr5 protein" (GenBank accession numbers NM_130773 (**Exhibit A**) and AB077881 (**Exhibit B**)). The caspr protein family is a well-studied protein family with known functional information, exactly the situation that Doerks *et al.* suggests will "simplify" and "avoid the pitfalls" of previous sequence-to-function methods of assigning protein function (Doerks *et al.*, page 248, columns 1 and 2). Thus, instead of supporting the Examiner's position against utility, Doerks *et al.* actually supports Appellants' position that the presently claimed sequences have a substantial and credible utility.

The Examiner next cited Smith *et al.* (Nature Biotechnology *15*:1222-1223, 1997) as teaching "that there are numerous cases in which proteins of very different functions are homologous" (the First Action at page 6). However, the Smith and Zhang article also states "the major problems associated with nearly all of the current automated annotation approaches are - paradoxically - minor database annotation inconsistencies (and a few outright errors)" (page 1222, second column, first paragraph, emphasis added). Thus, Smith and Zhang do not in fact seem to stand for the proposition that prediction of function based on homology is fraught with uncertainty, and thus also does not support the alleged lack of utility. The citation of Pilbeam *et al.* ("Pilbeam"; 1993, Bone 14:717-720), which allegedly details that "PTH and PTHrP are two structurally closely related proteins which can have opposite effects on bone resorption" (the First Action at page 6), is also hardly indicative of a high level of uncertainty in assigning function based on sequence. In fact, Pilbeam details that "the biological activities of hPTHrP 1-34 and synthetic bPTH 1-34 have generally been shown to be qualitatively similar" (Pilbeam at page 717), and thus also does not support the alleged lack of utility.

The Examiner next cited Brenner (TIG *15*:132-133, 1999) as teaching that "most homologs must have different molecular and cellular functions" (the First Action at page 6). However, this statement is based on the assumption that "if there are only 1000 superfamilies in nature, then most homologs must have different molecular and cellular functions" (Brenner, page 132, second column). Furthermore, Brenner suggests that one of the main problems in using homology to predict function is "an issue solvable by appropriate use of modern and accurate sequence comparison procedures" (Brenner, page 132, second column), and in fact references an article by Altschul *et al.*, which is the basis for one of the "modern and accurate sequence comparison procedures" used by Appellants. Thus, the Brenner article also does not support the alleged lack of utility.

The Examiner finally cited Bork *et al.* (Trends in Genetics *12*:425-427, 1996) as supporting the proposition that prediction of protein function from homology information is somewhat unpredictable, based on the "structural similarity of a small domain of the new protein to a small domain of a known protein" (the First Action at page 3). Thus, the Examiner's reliance on Bork *et al.* has the same failing as described above for Doerks *et al.*, specifically, the assumption that Appellants assertion that the present sequences are caspr proteins are made on the basis of structural similarity of a small domain of the new protein to a small domain of a known protein. Appellants again would like to invite the Board's attention to the fact that two sequences sharing nearly 100% percent identity at the protein level with the claimed sequence are present in the leading scientific repository for biological sequence data (GenBank), and have been annotated by third party scientists *wholly unaffiliated with Appellants* as "Homo sapiens caspr5 protein" (GenBank accession numbers NM_130773 (**Exhibit A**) and AB077881 (**Exhibit B**)). Thus, Appellants assertion that the present sequences are caspr proteins are not made on the basis of "structural similarity of a small domain of the new protein to a small domain of a known protein", but rather vast homology over the entire sequence. Thus, Bork *et al.* also does not support the alleged lack of utility for the present invention.

Thus, while Appellants have provided evidence of record that conclusively establishes that those skilled in the art would believe that the specifically claimed sequence encodes a caspr protein, the Examiner has provided no evidence that directly establishes that the specifically claimed sequence does not encode a caspr protein. Accordingly, the evidence of record compels a finding that the present invention has a patentable utility.

Furthermore, with regard to the citation of journal articles to support an allegation of a lack of utility, the PTO has repeatedly attempted to deny the utility of nucleic acid sequences based on a small number of publications that call into doubt prediction of protein function from homology information and the usefulness of bioinformatic predictions, of which these articles are merely the latest examples. Appellants readily agree that there is not 100% consensus within the scientific community regarding prediction of protein function from homology information, and further agree that prediction of protein function from homology information is not 100% accurate. However, Appellants respectfully point out that the lack of 100% consensus on prediction of protein function from homology information is **completely irrelevant** to the question of whether the claimed nucleic acid sequence has a substantial and specific utility, and that 100% accuracy of prediction of protein function from homology information is **not the standard** for patentability under 35 U.S.C. § 101. Appellants respectfully point out that, as discussed above, the legal test for utility simply involves an assessment of whether those skilled in the art would find any of the utilities described for the invention to be **believable**. Appellants submit that the overwhelming majority of those of skill in the relevant art would **believe** prediction of protein function from homology information and the usefulness of bioinformatic predictions to be powerful and useful tools, as evidenced by hundreds if not thousands of journal articles (which Appellants will submit to the Office if the Board truly doubts Appellants' assertion that the overwhelming majority of those of skill in the art place a high value on prediction of protein function from homology information and the usefulness of bioinformatic predictions), and would thus **believe** that Appellants sequence is a caspr protein. As **believability** is the standard for meeting the utility requirement of 35 U.S.C. § 101, and **not** 100% consensus or 100% accuracy, Appellants submit that the present claims must clearly meet the requirements of 35 U.S.C. § 101.

Furthermore, the **PTO itself** does not require 100% identity between proteins to establish functional homology. Example 10 of the Revised Interim Utility Guidelines Training Materials, discussed above, only requires a similarity score greater than 95% to establish functional homology. Thus, scientific publications that generally assert that very small changes between amino acid sequences can lead to changes in function, or publications describing specific examples of proteins, distinct from Appellants sequence, where a minor change in amino acid sequence has lead to a change in function, have been viewed by the PTO itself as irrelevant to the question of utility, and thus do not support the Examiner's allegation that the presently claimed sequence lacks utility. Therefore, the present utility rejection must fail as a matter of policy, as a matter of science, and as a matter of law.

However, although Appellants need only make one credible assertion of utility to meet the requirements of 35 U.S.C. § 101 (*Raytheon v. Roper*, 220 USPQ 592 (Fed. Cir. 1983); *In re Gottlieb*, 140 USPQ 665 (CCPA 1964); *In re Malachowski*, 189 USPQ 432 (CCPA 1976); *Hoffman v. Klaus*, 9 USPQ2d 1657 (Bd. Pat. App. & Inter. 1988)), as yet another example of the utility of the present sequence, Appellants pointed out both in the Response to the First Action and in the Response to the Final Action that the present nucleic acid sequences have utility in forensic analysis (see, for example, the specification at page 10, lines 15-19). As described in the specification at page 15, lines 21-25, the present sequences define a coding single nucleotide polymorphism - specifically, a C/T polymorphism at position 812 of SEQ ID NO:1, which can lead to a serine or leucine residue at amino acid position 271 of SEQ ID NO:2. As such polymorphisms are the basis for forensic analysis, which in undoubtedly a "real world" utility, the present sequences must in themselves be useful.

In the Final Action, the Examiner questioned this asserted utility, stating that "any polynucleotide containing a SNP can be used for diagnostic assays" (the Final Action at page 3). The Examiner seems to be confusing the requirements of a **specific** utility with a **unique** utility. The fact that other polymorphic markers have been identified in other genetic loci, or that the use of the presently described polymorphic markers will provide additional information concerning the prevalence of these markers in certain subpopulations, does not mean that use of the polymorphic markers identified by Appellants' in SEQ ID NO:1 is not a specific utility. As clearly stated by the Federal Circuit in *Carl Zeiss Stiftung v. Renishaw*

PLC, 20 USPQ2d 1101 (Fed. Cir. 1991):

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding a lack of utility." *Envirotech Corp. v. Al George, Inc.*, 221 USPQ 473, 480 (Fed. Cir. 1984)

In other words, just because other (possibly better) polymorphic markers from the human genome have been described, or that additional information about the presently described polymorphic markers can be gained through the use of these markers, does not establish that the presently described polymorphic markers lack a specific utility. The requirement for a **specific** utility, which is part of the standard for utility under 35 U.S.C. § 101 presently being applied by the Office, should not be confused with the requirement for a **unique** utility, which is not the legal standard. If every invention were required to have a unique utility, the Patent and Trademark Office would no longer be issuing patents on batteries, automobile tires, golf balls, golf clubs, and treatments for a variety of human diseases, just to name a few particular examples, because other examples of each of these have already been described and patented. However, only the briefest perusal of virtually any issue of the Official Gazette provides numerous examples of patents being granted on each of the above compositions every week. Furthermore, if each invention needed to have a unique utility in order to be patented, the entire class and subclass system would be an effort in futility, as the class and subclass system serves solely to group such common inventions, which would not be required if each invention needed to have a unique utility. In view of the above standards and "common sense" analysis, there can be little question that the present sequence clearly meets the requirements of 35 U.S.C. § 101.

The Final Action states that "without knowing the functions (i.e. utility) of the polynucleotide and protein of the present invention, one cannot assess a utility for the diagnostic assays using these molecules" (the Final Action at page 3). Appellants respectfully submit that the Examiner has **completely** missed Appellants point with regard to the use of the presently described polymorphism in **forensic** analysis. In forensic analysis, polymorphic markers, such as the presently described polymorphism, can be used by those skilled in the art to distinguish one person from another based on the presence or absence of the described polymorphism. Forensic analysis requires **no** information regarding the function of the protein encoded by the polymorphic DNA sequence. The Examiner has provided no evidence of record that establishes that skilled artisans would not be able to use the presently described polymorphism in forensic analysis exactly as it was described in the specification as originally filed, without any additional research. It is important to note that simply because the use of this polymorphic marker will necessarily provide additional information on the percentage of particular subpopulations that contain this polymorphic marker does not mean that additional research is needed in order for this marker as it is presently described in the instant specification to be used in forensic science. Thus, the Examiner has failed to meet his evidentiary burden of proving that the present invention lacks utility.

This is also not a case of a probable utility. Appellants point out that even in the worst case scenario, the described polymorphism is useful to distinguish 50% of the population (in other words, the marker being present in half of the population). Appellants point out that the ability of a polymorphic marker to distinguish at least 50% of the population is an inherent feature of any polymorphic marker, and this feature is well understood by those of skill in the art. Appellants note that as a matter of law, it is well settled that a patent need not disclose what is well known in the art. *In re Wands*, *supra*. Appellants respectfully point out that all that is required to support Appellants assertion of utility is for the skilled artisan to believe that the presently described polymorphic marker could be useful in forensic analysis. The fact that forensic biologists use polymorphic markers such as that described by Appellants every day provides more that ample support for the assertion that forensic biologists would also be able to use the specific polymorphic marker described by Appellants in the same fashion. Therefore, these allegations are completely without merit, and in no way establish that the present invention lacks utility.

Further, as the presently described polymorphisms are part of the family of polymorphisms that have a well established utility, Appellants reliance on *In re Brana*, (34 USPQ2d 1436 (Fed. Cir. 1995), "*Brana*") is directly on point. In *Brana*, the Federal Circuit admonished the Patent and Trademark Office for confusing "the requirements under the law for obtaining a patent with the requirements for obtaining government approval to market a particular drug for human consumption". *Brana* at 1442. The Federal Circuit went on to state:

> At issue in this case is an important question of the legal constraints on patent office examination practice and policy. The question is, with regard to pharmaceutical inventions,

> what must the applicant provide regarding the practical utility or usefulness of the invention for which patent protection is sought. This is not a new issue; it is one which we would have thought had been settled by case law years ago.

Brana at 1439, emphasis added. The choice of the phrase "utility or usefulness" in the foregoing quotation is highly pertinent. The Federal Circuit is evidently using "utility" to refer to rejections under 35 U.S.C. § 101, and is using "usefulness" to refer to rejections under 35 U.S.C. § 112, first paragraph. This is made evident in the continuing text in *Brana*, which explains the correlation between 35 U.S.C. §§ 101 and 112, first paragraph. The Federal Circuit concluded:

> FDA approval, however, is not a prerequisite for finding a compound useful within the meaning of the patent laws. Usefulness in patent law, and in particular in the context of pharmaceutical inventions, necessarily includes the expectation of further research and development. The stage at which an invention in this field becomes useful is well before it is ready to be administered to humans. Were we to require Phase II testing in order to prove utility, the associated costs would prevent many companies from obtaining patent protection on promising new inventions, thereby eliminating an incentive to pursue, through research and development, potential cures in many crucial areas such as the treatment of cancer.

Brana at 1442-1443, citations omitted, emphasis added. As set forth above, the present polymorphism is useful in forensic analysis exactly as it is described in the specification as originally filed, without the need for any further research. However, even if, *arguendo*, further research might be required in certain aspects of the present invention, this does not preclude a finding that the invention has utility, as set forth by the Federal Circuit's holding in *Brana*, which clearly states, as highlighted in the quote above, that "pharmaceutical inventions, necessarily includes the expectation of further research and development" (*Brana* at 1442-1443, emphasis added). In assessing the question of whether undue experimentation would be required in order to practice the claimed invention, the key term is "undue", not "experimentation". *In re Angstadt and Griffin*, 190 USPQ 214 (CCPA 1976). The need for some experimentation does not render the claimed invention unpatentable. Indeed, a considerable amount of experimentation may be permissible if such experimentation is routinely practiced in the art. *In re Angstadt and Griffin, supra*; *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 18 USPQ2d 1016 (Fed. Cir. 1991). As a matter of law, it is well settled that a patent need not disclose what is well known in the art.

In re Wands, supra.

As yet another example of the utility of the present sequence, Appellants pointed out in the response to the First Action that those of skill in the art would readily appreciate the importance of tracking the expression of the gene encoding the described protein, as described in the specification as originally filed, at least at page 5, lines 12-14. In particular, the specification describes how the described sequences can be represented using a gene chip format to provide a high throughput analysis of the level of gene expression. Such "DNA chips" clearly have utility, as evidenced by hundreds of issued U.S. Patents, as exemplified by U.S. Patent Nos. 5,445,934 (**Exhibit W**), 5,556,752 (**Exhibit X**), 5,744,305 (**Exhibit Y**), 5,837,832 (**Exhibit Z**), 6,156,501 (**Exhibit AA**) and 6,261,776 (**Exhibit BB**). Appellants point out that expression profiling does not require a knowledge of the function of the particular nucleic acid on the chip - rather the gene chip indicates which DNA fragments are expressed at greater or lesser levels in two or more particular tissue types.

Evidence of the "real world" <u>substantial</u> utility of the present invention is further provided by the fact that there is an entire industry established based on the use of gene sequences or fragments thereof in a gene chip format. Perhaps the most notable gene chip company is Affymetrix. However, there are many companies that have, at one time or another, concentrated on the use of gene sequences or fragments, in gene chip and non-gene chip formats, for example: Gene Logic, ABI-Perkin-Elmer, HySeq and Incyte. In addition, one such company (Rosetta Inpharmatics) was viewed to have such "real world" value that it was acquired by large a pharmaceutical company (Merck) for significant sums of money (net equity value of the transaction was $620 million). The "real world" <u>substantial</u> industrial utility of gene sequences or fragments would, therefore, appear to be widespread and well established. Clearly, there can be no doubt that the skilled artisan would know how to use the presently claimed sequences (see Section VIII(B), below), strongly arguing that the claimed sequences have utility. Given the widespread utility of such "gene chip" methods using *public domain* gene sequence information, there can be little doubt that the use of the presently described *novel* sequences would have great utility in such DNA chip applications. As the present sequences are <u>specific</u> markers of the human genome (see below), and such <u>specific</u> markers are targets for the discovery of drugs that are associated with human disease, as described above, those of skill in the art would instantly recognize that the present nucleotide sequences would be ideal, novel candidates for assessing gene expression using such DNA chips. Clearly, compositions that enhance the utility of such DNA chips, such as the presently claimed nucleotide sequences, must in themselves be useful. Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

The Examiner also dismisses this assertion of utility, stating that "any nucleotide sequence can be used in such an assay" (the Final Action at page 4). Appellants first point out that the present sequence, which has been biologically validated to be expressed, has a much greater utility than sequences that are merely predicted to be expressed based on bioinformatic analysis. Second, not "any nucleotide sequence" can be used to track gene expression, but rather, only those small percentage of nucleotide sequences that are expressed can be used in such a manner. Third, the Examiner again seems to be confusing the requirements of a **specific** utility with a **unique** utility. The fact that other nucleotide sequences can be used to track gene expression does not mean that the use of Appellants' sequence to track gene expression is not a specific utility (*Carl Zeiss Stiftung v. Renishaw PLC, supra*). Therefore, this argument completely fails to support the alleged lack of utility of the presently claimed compositions.

Clearly, persons of skill in the art, as well as venture capitalists and investors, readily recognize the utility, both scientific and commercial, of genomic data in general, and specifically human genomic data. Billions of dollars have been invested in the human genome project, resulting in useful genomic data (see, *e.g.*, Venter *et al.*, 2001, Science *291*:1304; **Exhibit CC**). The results have been a stunning success as the utility of human genomic data has been widely recognized as a great gift to humanity (see, *e.g.*, Jasny and Kennedy, 2001, Science *291*:1153; **Exhibit DD**). Clearly, the usefulness of human genomic data, such as the presently claimed nucleic acid molecules, is substantial and credible (worthy of billions of dollars and the creation of numerous companies focused on such information) and well-established (the utility of human genomic information has been clearly understood for many years).

Additionally, as set forth by Appellants in the response to the First Action, and as described in the specification as originally filed, at least at page 10, line 18, the present nucleotide sequence has a specific utility in determining the genomic structure of the corresponding human chromosome, for example mapping the protein encoding regions. The claimed polynucleotide sequences defining how the encoded exons are actually spliced together to produce an active transcript (*i.e.*, the described sequences are useful for functionally defining exon splice-junctions). This is evidenced by the fact that SEQ ID NO:1 can be used to map the 24 coding exons on chromosome 2 (present within six overlapping chromosome 2 clones; GenBank Accession Numbers AC097715, AC019105, AC019159, AC104648, AC074362 and 079154 alignments and the first page from the GenBank reports are presented in **Exhibit EE**). In disclosing biologically validated exon splice junctions, the claimed sequence provides physical evidence that effectively trumps the hypothetical conclusions provided by bioinformatics analysis of the corresponding genomic region conducted without supporting physical data. Thus, the claimed sequence clearly meet the requirements of 35 U.S.C. § 101.

Appellants respectfully remind the Board that only a minor percentage of the genome (2-4%) actually encodes exons, which in-turn encode amino acid sequences. The presently claimed polynucleotide sequence provides biologically validated empirical data (*e.g.,* showing which sequences are transcribed, spliced, and polyadenylated) that *specifically* define that portion of the corresponding genomic locus that actually encodes exon sequence. Appellants respectfully submit that the practical scientific value of expressed, spliced, and polyadenylated mRNA sequences is readily apparent to those skilled in the relevant biological and biochemical arts. The specification details that "sequences derived from regions adjacent to the intron/exon boundaries of the human gene can be used to design primers for use in amplification assays to detect mutations within the exons, introns, splice sites (*e.g.*, splice acceptor and/or donor sites), *etc.*, that can be used in diagnostics and pharmacogenomics" (specification at page 10, lines 19-24). Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Thus, as set forth in the Response to the First Action, the present nucleotide sequence has a specific utility in mapping the claimed sequence to the corresponding human chromosome, specifically chromosome 2, as described above. Clearly, the present polynucleotide provides exquisite specificity in localizing the specific region of human chromosome 2 that contains the gene encoding the given polynucleotide, a utility not shared by virtually any other nucleic acid sequences. In fact, it is this specificity that makes this particular sequence so useful. Early gene mapping techniques relied on methods such as Giemsa staining to identify regions of chromosomes. However, such techniques produced genetic maps with a resolution of only 5 to 10 megabases, far too low to be of much help in identifying specific genes involved in disease. The skilled artisan readily appreciates the significant benefit afforded by markers that map a specific locus of the human genome, such as the present nucleic acid sequence. For further evidence in support of the Appellants' position, the Board is requested to review, for example, section 3 of Venter *et al.* (*supra*, at pp. 1317-1321, including Fig. 11 at pp.1324-1325; see **Exhibit CC**), which demonstrates the significance of expressed sequence information in the structural analysis of genomic data. The presently claimed polynucleotide sequence defines a biologically validated sequence that provides a unique and specific resource for mapping the genome essentially as described in the Venter *et al.* article Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

The Examiner also questions this asserted utility, stating that "any nucleotide sequence can be used in such an assay" (the Final Action at page 4). Appellants first point out that only those small percentage of nucleotide sequences that are located in this region of chromosome 2 can be used in such a manner. Second, the Examiner again seems to be confusing the requirements of a **specific** utility with a **unique** utility. The fact that a small number of other nucleotide sequences could be used to map the protein coding regions in this specific region of chromosome 2 does not mean that the use of Appellants' sequence to map the protein coding regions of chromosome 2 is not specific (*Carl Zeiss Stiftung v. Renishaw PLC, supra*).

Importantly, it has been clearly established that a statement of utility in a specification must be accepted absent reasons why one skilled in the art would have reason to doubt the objective truth of such statement. *In re Langer*, 503 F.2d 1380, 1391, 183 USPQ 288, 297 (CCPA, 1974; "*Langer*"); *In re Marzocchi*, 439 F.2d 220, 224, 169 USPQ 367, 370 (CCPA, 1971). As clearly set forth in *Langer*:

> As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope.

Langer at 297, emphasis in original. As set forth in the MPEP, "Office personnel must provide evidence sufficient to show that the statement of asserted utility would be considered 'false' by a person of ordinary skill in the art" (MPEP, Eighth Edition at 2100-40, emphasis added). Thus, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Regarding the utility requirements under 35 U.S.C. § 101, the Federal Circuit has clearly stated "(t)he threshold of utility is not high: An invention is 'useful' under section 101 if it is capable of providing some identifiable benefit." *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d 1364, 51 USPQ2d 1700 (Fed. Cir. 1999) (citing *Brenner v. Manson*, 383 U.S. 519, 534 (1966)). Additionally, the Federal Circuit has stated that "(t)o violate § 101 the claimed device must be totally incapable of achieving a useful result." *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571, 24 USPQ2d 1401 (Fed. Cir. 1992), emphasis added. *Cross v. Iizuka* (753 F.2d 1040, 224 USPQ 739 (Fed. Cir. 1985); "*Cross*") states "any utility of the claimed compounds is sufficient to satisfy 35 U.S.C. § 101". *Cross* at 748, emphasis added. Indeed, the Federal Circuit recently emphatically confirmed that "anything under the sun that is made by man" is patentable (*State Street Bank & Trust Co. v. Signature Financial Group Inc.*, 149 F.3d 1368, 47 USPQ2d 1596, 1600 (Fed. Cir. 1998), citing the U.S. Supreme Court's decision in *Diamond vs. Chakrabarty*, 447 U.S. 303, 206 USPQ 193 (U.S., 1980)). Thus, based on the relevant case law, the present claims clearly meet the requirements of 35 U.S.C. § 101.

Finally, While Appellants are well aware of the new Utility Guidelines set forth by the USPTO, Appellants respectfully point out that the current rules and regulations regarding the examination of patent applications is and always has been the patent laws as set forth in 35 U.S.C. and the patent rules as set forth in 37 C.F.R., not the Manual of Patent Examination Procedure or particular guidelines for patent examination set forth by the USPTO. Furthermore, it is the job of the judiciary, not the USPTO, to interpret these laws and rules. Appellants are unaware of any significant recent changes in either 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit that is in keeping with the new Utility Guidelines set forth by the USPTO. This is underscored by numerous patents that have been issued over the years that claim nucleic acid fragments that do not comply with the new Utility Guidelines. As examples of such issued U.S. Patents, the Board is invited to review U.S. Patent Nos. 5,817,479 (**Exhibit FF**), 5,654,173 (**Exhibit GG**), and 5,552,281 (**Exhibit HH**; each of which claims short polynucleotides), and recently issued U.S. Patent No. 6,340,583 (**Exhibit II**; which includes no working examples), none of which contain examples of the "real-world" utilities that the Examiner seems to be requiring. As issued U.S. Patents are presumed to meet all of the requirements for patentability, including 35 U.S.C. §§ 101 and 112, first paragraph (see Section VIII(B), below), Appellants submit that the present polynucleotides must also meet the requirements of 35 U.S.C. § 101. While Appellants agree that each application is examined on its own merits, Appellants are unaware of any changes to 35 U.S.C. § 101, or in the interpretation of 35 U.S.C. § 101 by the Supreme Court or the Federal Circuit, since the issuance of these patents that render the subject matter claimed in these patents, which is similar to the subject matter in question in the present application, as suddenly non-statutory or failing to meet the requirements of 35 U.S.C. § 101. Thus, holding Appellants to a different standard of utility would be arbitrary and capricious, and, like other clear violations of due process, cannot stand.

For each of the foregoing reasons, Appellants submit that the rejection of claims 1-3 and 6-8 under 35 U.S.C. § 101 must be overruled.

B. Are Claims 1-3 and 6-8 Unusable Due to a Lack of Patentable Utility?

The Final Action next rejects claims 1-3 under 35 U.S.C. § 112, first paragraph, since allegedly one skilled in the art would not know how to use the invention, as the invention allegedly is not supported by either a clear asserted utility or a well-established utility.

The arguments detailed above in Section VIII(A) concerning the utility of the presently claimed sequences are incorporated herein by reference. As the Federal Circuit and its predecessor have determined that the utility requirement of Section 101 and the how to use requirement of Section 112, first paragraph, have the same basis, specifically the disclosure of a credible utility (*In re Brana*, *supra*; *In re Jolles*, 628 F.2d 1322, 1326 n.11, 206 USPQ 885, 889 n.11 (CCPA 1980); *In re Fouche*, 439 F.2d 1237, 1243, 169 USPQ 429, 434 (CCPA 1971)), Appellants submit that as claims 1-3 and 6-8 have been shown to have "a specific, substantial, and credible utility", as detailed in Section VIII(A) above, the present rejection of claims 1-3 and 6-8 under 35 U.S.C. § 112, first paragraph, cannot stand.

Appellants therefore submit that the rejection of claims 1-3 and 6-8 under 35 U.S.C. § 112, first paragraph, must be overruled.

IX. APPENDIX

The claims involved in this appeal are as follows:

1. (Amended) An isolated nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO:1.

2. (Amended) An isolated nucleic acid molecule comprising a nucleotide sequence that:

   (a) encodes the amino acid sequence of SEQ ID NO:2; and

   (b) hybridizes to the complement of the nucleotide sequence of SEQ ID NO:1 under highly stringent conditions of 0.5 M $NaHPO_4$, 7% sodium dodecyl sulfate (SDS) and 1 mM EDTA at 65°C and washing in 0.1x SSC/0.1%SDS at 68°C.

3. An isolated nucleic acid molecule comprising a nucleotide sequence that encodes the amino acid sequence shown in SEQ ID NO:2.

6. A recombinant expression vector comprising the isolated nucleic acid molecule of claim 3.

7. The recombinant expression vector of claim 6, wherein the nucleic acid molecule comprises the nucleotide sequence of SEQ ID NO:1.

8. A host cell comprising the recombinant expression vector of claim 6.

X. CONCLUSION

Appellants respectfully submit that, in light of the foregoing arguments, the Final Action's conclusion that claims 1-3 and 6-8 lack a patentable utility and are unusable by the skilled artisan due to a lack of patentable utility, are unwarranted. It is therefore requested that the Board overturn the Final Action's rejections.

Respectfully submitted,

July 28, 2003
Date

David W. Hibler

David W. Hibler Reg. No. 41,071
Agent For Appellants

LEXICON GENETICS INCORPORATED
8800 Technology Forest Place
The Woodlands, TX 77381
(281) 863-3399

24231
PATENT TRADEMARK OFFICE

TABLE OF AUTHORITIES

CASES

STATUTES

A

```
>NM_130773 ACCESSION:NM_130773 NID: gi 20544138 ref NM_130773.2  Homo
          sapiens caspr5 protein (caspr5), transcript variant 1,
          mRNA
        Length = 5284

 Score = 2567 bits (6581), Expect = 0.0
 Identities = 1303/1307 (99%), Positives = 1303/1307 (99%), Gaps = 3/1307 (0%)
 Frame = +2

Query: 1    MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSPAQLNW 60
             MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSPAQLNW
Sbjct: 365  MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSPAQLNW 544

Query: 61   RVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRNWKQY 120
             RVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRNWKQY
Sbjct: 545  RVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRNWKQY 724

Query: 121  KQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDV 180
             KQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDV
Sbjct: 725  KQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDV 904

Query: 181  ADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKGRLAL 240
             ADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKGRLAL
Sbjct: 905  ADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKGRLAL 1084

Query: 241  HLNLGDSKARLSSSLPSATLGSLLDDQHWH-VLIERVGKQVNFTVDKHTQHFRTKGETDA 299
             HLNLGDSKARLSSSLPSATLGSLLDDQHWH VLIERVGKQVNFTVDKHTQHFRTKGETDA
Sbjct: 1085 HLNLGDSKARLSSSLPSATLGSLLDDQHWHSVLIERVGKQVNFTVDKHTQHFRTKGETDA 1264

Query: 300  LDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNII-LAKRRKHQIYTVGNVTFS 358
             LDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNII LAKRRKHQIYT GNVTFS
Sbjct: 1265 LDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNIIDLAKRRKHQIYT-GNVTFS 1441

Query: 359  CSEPQIVPITF-NSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTLLLS 417
             CSEPQIVPITF NSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTLLLS
Sbjct: 1442 CSEPQIVPITFVNSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTLLLS 1621

Query: 418  LEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPDSTW 477
             LEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPDSTW
Sbjct: 1622 LEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPDSTW 1801

Query: 478  VQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNFSDL 537
             VQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNFSDL
Sbjct: 1802 VQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNFSDL 1981

Query: 538  HIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGN 597
             HIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGN
Sbjct: 1982 HIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGN 2161

Query: 598  TAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGG 657
             TAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGG
Sbjct: 2162 TAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGG 2341

Query: 658  SMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQ 717
             SMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQ
Sbjct: 2342 SMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQ 2521
```

```
Query: 718  CECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIG 777
            CECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIG
Sbjct: 2522 CECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIG 2701

Query: 778  PLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIR 837
            PLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIR
Sbjct: 2702 PLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIR 2881

Query: 838  LEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRE 897
            LEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRE
Sbjct: 2882 LEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRE 3061

Query: 898  TSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGH 957
            TSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGH
Sbjct: 3062 TSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGH 3241

Query: 958  CSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVT 1017
            CSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVT
Sbjct: 3242 CSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVT 3421

Query: 1018 KNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYH 1077
            KNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYH
Sbjct: 3422 KNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYH 3601

Query: 1078 LNKEETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLT 1137
            LNKEETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLT
Sbjct: 3602 LNKEETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLT 3781

Query: 1138 LGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSC 1197
            LGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSC
Sbjct: 3782 LGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSC 3961

Query: 1198 GFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMT 1257
            GFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMT
Sbjct: 3962 GFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMT 4141

Query: 1258 RFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI 1304
            RFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI
Sbjct: 4142 RFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI 4282
```

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Boo

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display default Save Text Add to Clipboard Get Subsequence

□ 1: NM_130773. Homo sapiens casp...[gi:20544138] Links

```
LOCUS       caspr5                  5284 bp    mRNA    linear   PRI 05-NOV-2002
DEFINITION  Homo sapiens caspr5 protein (caspr5), transcript variant 1, mRNA.
ACCESSION   NM_130773
VERSION     NM_130773.2  GI:20544138
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Takeuchi,K., Watanabe,N., Kawano,T. and Kawamura,K.
  TITLE     In vitro and in vivo studies on the involvement of neural cell
            adhesion molecules and chondroitin sulfate proteoglycans in
            defining discrete axonal pathways of the rat cerebral cortex
  JOURNAL   Unpublished
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AB077881.1 and AK056528.1.
            On May 13, 2002 this sequence version replaced gi:18640733.
            Summary: This gene product belongs to the neurexin family, members
            of which function in the vertebrate nervous system as cell adhesion
            molecules and receptors. This protein, like other neurexin
            proteins, contains epidermal growth factor repeats and laminin G
            domains. In addition, it includes an F5/8 type C domain,
            discoidin/neuropilin- and fibrinogen-like domains, and
            thrombospondin N-terminal-like domains. Alternative splicing of
            this gene results in 2 transcript variants encoding different
            isoforms.
            Transcript Variant: This variant (1) encodes the longer isoform
            (1).
FEATURES             Location/Qualifiers
     source          1..5284
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="2"
                     /map="2q14.1"
     gene            1..5284
                     /gene="caspr5"
                     /note="synonym: FLJ31966"
                     /db_xref="LocusID:129684"
     CDS             365..4285
                     /gene="caspr5"
                     /codon_start=1
                     /product="caspr5 protein isoform 1"
                     /protein_id="NP_570129.1"
                     /db_xref="GI:18640734"
                     /db_xref="LocusID:129684"
                     /translation="MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFS
                     SSSDLTGTHSPAQLNWRVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDW
                     VTSYSLMFSDTGRNWKQYKQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWN
                     PSGKIGMRVEVYGCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVL
                     FHGEGQRGDHITLELQKGRLALHLNLGDSKARLSSSLPSATLGSLLDDQHWHSVLIER
                     VGKQVNFTVDKHTQHFRTKGETDALDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLY
                     YNGVNIIDLAKRRKHQIYTGNVTFSCSEPQIVPITFVNSSGSYLLLPGTPQIDGLSVS
                     FQFRTWNKDGLLLSTELSEGSGTLLLSLEGGILRLVIQKMTERVAEILTGSNLNDGLW
                     HSVSINARRNRITLTLDDEAAPPAPDSTWVQIYSGNSYYFGGCPDNLTDSQCLNPIKA
                     FQGCMRLIFIDNQPKDLISVQQGSLGNFSDLHIDLCSIKDRCLPNYCEHGGSCSQSWT
                     TFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGNTAGFFYIDSDGSGPLGPLQVYCNIT
                     EDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGGSMEQLEAVIDGSEHCEQEVAYHC
                     RRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQCECGLDESCLDIQHFCNCDAD
                     KDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIGPLRCYGDRRFWNAVSFYTE
                     ASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIRLEISSPSEITFAIDVGN
                     GPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRETSEEGHFRLQLNSQL
                     FVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGHCSSYGSICHNGGK
                     CVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVTKNISLSSSAIY
                     TDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYHLNKEETHVF
                     TIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLTLGKVTEN
                     LGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSCGFMVD
                     SDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMTRFL
                     YQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI"
     misc_feature    530..877
                     /gene="caspr5"
                     /note="F5_F8_type_C; Region: F5/8 type C domain. This
                     domain is also known as the discoidin (DS) domain family.
                     The bacterial examples are not yet included in the SEED
                     alignment and are only found with low scores"
                     /db_xref="CDD:pfam00754"
     misc_feature    551..886
                     /gene="caspr5"
                     /note="FA58C; Region: Coagulation factor 5/8 C-terminal
                     domain, discoidin domain"
                     /db_xref="CDD:smart00231"
     misc_feature    977..1378
                     /gene="caspr5"
                     /note="LamG; Region: Laminin G domain"
                     /db_xref="CDD:smart00282"
     misc_feature    989..1384
                     /gene="caspr5"
                     /note="laminin_G; Region: Laminin G domain"
                     /db_xref="CDD:pfam00054"
     misc_feature    1529..1927
                     /gene="caspr5"
                     /note="LamG; Region: Laminin G domain"
                     /db_xref="CDD:smart00282"
     misc_feature    1688..1936
                     /gene="caspr5"
                     /note="laminin_G; Region: Laminin G domain"
                     /db_xref="CDD:pfam00054"
     misc_feature    2798..3178
                     /gene="caspr5"
                     /note="LamG; Region: Laminin G domain"
                     /db_xref="CDD:smart00282"
     misc_feature    2819..3187
                     /gene="caspr5"
                     /note="laminin_G; Region: Laminin G domain"
                     /db_xref="CDD:pfam00054"
     misc_feature    3263..3346
                     /gene="caspr5"
                     /note="EGF; Region: EGF-like domain. There is no clear
                     separation between noise and signal. pfam00053 is very
                     similar, but has 8 instead of 6 conserved cysteines.
                     Includes some cytokine receptors. The family is difficult
                     to model due to many similar but different sub-types of
                     EGF domains"
                     /db_xref="CDD:pfam00008"
     misc_feature    3470..3880
                     /gene="caspr5"
                     /note="LamG; Region: Laminin G domain"
                     /db_xref="CDD:smart00282"
     misc_feature    3503..3880
                     /gene="caspr5"
                     /note="laminin_G; Region: Laminin G domain"
                     /db_xref="CDD:pfam00054"
     polyA_signal    5271..5276
                     /gene="caspr5"
BASE COUNT     1364 a   1314 c   1305 g   1301 t
ORIGIN
        1 gattccagct ctcgcgcccg acgaggtgga tttggctgtc caccgagctc cggcgcctgt
       61 cgttctaatt gggtttggat ttgcaccgtt aaggaggggg gaagagaagg aagaggcggg
      121 cgaggaaggc gagtccagct agcggctgtt gcggggaccg tagccccagc tgcagctccg
      181 aagaatcccc cgccacggtt tcggtggagc gtctgggcac gggatggagt gaaagagcga
      241 gtgcctctcc aagcgggggt gggagggggt caggctgtgc agaggagaga gacagcgaga
      301 agaagccgcg gctggctact gcgaatttgg gattcgattg ggagggaccg ctcactcggg
      361 ggaaatggat tctttaccac ggctgaccag cgttttgact ttgctgttct ctggcttgtg
      421 gcatttagga ttaacagcga caaactacaa ctgtgatgat ccactagcat ccctgctctc
      481 tccaatggct ttttccagtt cctcagacct cactggcact cacagcccag ctcaactcaa
      541 ctggagagtt ggaactggcg gttggtcccc agcagattcc aatgctcaac agtggctcca
      601 gatggacctg ggaaacagag tagagattac agcagtggcc acgcagggaa gatacggaag
      661 ctctgactgg gtgacgagtt acagcctgat gttcagtgac acaggacgca actggaaaca
      721 gtacaaacaa gaagacagca tctggacctt tgcaggaaac atgaatgctg acagcgtggt
      781 gcaccacaag ctattgcact cagtgagagc ccgatttgtt cgctttgtgc ccctggaatg
      841 gaatcccagt gggaagattg gcatgagagt cgaggtctac ggatgttcct ataaatcaga
      901 tgttgctgac tttgatggcc gaagctcact tctgtacagg ttcaatcaga agttgatgag
      961 tactctcaaa gatgtgatct ccctgaagtt caagagcatg caaggagatg ggtcctgtt
     1021 ccatggagaa ggtcagcgtg gagaccacat caccttggaa ctccagaagg ggaggctcgc
     1081 cctacacctc aatttgggtg acagcaaagc gcggctcagc agcagcttgc cctctgccac
     1141 cctgggcagc ctcctggatg accagcactg gcactcggtc ctcattgagc gggtgggcaa
     1201 gcaggtgaac ttcacggtgg acaagcacac acagcacttc cgcaccaagg gcgagacgga
     1261 tgccttagac attgactatg agcttagttt tggaggaatt ccagtaccag gaaaacctgg
     1321 gaccttttta aagaaaaact tccatggatg catcgaaaac ctttactaca atggagtaaa
     1381 cataattgac ctggctaaga gacgaaagca tcagatctat actggcaatg tcactttttc
     1441 ctgctccgaa ccacagattg tgcccatcac atttgtcaac tccagcggca gctatttgct
     1501 gctgcccggc accccccaaa ttgatgggct ctcagtgagt ttccagtttc gaacatggaa
     1561 caaggatggt ctgcttctgt ccacagagct gtctgagggc tcgggaaccc tgctgctgag
     1621 cctggagggt ggaatcctga gactcgtgat tcagaaaatg acagaacgcg tagctgaaat
     1681 cctcacaggc agcaacttga atgatggcct gtggcactcg gttagcatca acgccaggag
     1741 gaaccgcatc acgctcactc tggatgatga agcagcaccc ccggctccag acagcacttg
     1801 ggtgcagatt tattctggaa atagctacta ttttggaggg tgccccgaca atctcaccga
     1861 ttcccaatgt ttaaatccca ttaaggcttt ccaaggctgc atgaggctca tctttattga
     1921 taaccagccc aaggacctca tttcagttca gcaaggttcc ctggggaatt ttagtgattt
     1981 acacattgat ctgtgtagca tcaaagacag gtgtttgcca aactactgtg aacatggagg
     2041 aagctgctcc cagtcctgga ctaccttcta ttgtaactgc agtgacacaa gttacactgg
     2101 tgccacctgc cacaactcca tctacgagca atcctgcgag gtgtacaggc accaggggaa
     2161 tacagccggc ttcttctaca tcgactcaga tggcagcggc ccactgggac ctctccaggt
     2221 gtactgcaat atcactgagg acaagatctg gacatcagtg cagcacaaca atacagagct
     2281 gacccgagtg cggggcgcta accctgagaa gccctatgcc atggccttgg actacggggg
     2341 cagcatggaa cagctggagg ccgtgatcga cggctctgag cactgtgagc aggaggtggc
     2401 ctaccactgc aggaggtccc gcctgctcaa cacgccggat ggaacaccat ttacctggtg
     2461 gattgggcgg tccaatgaaa ggcaccctta ctggggaggt tcccctcctg gggtccagca
     2521 gtgtgagtgt ggcctagacg agagctgcct ggacattcag cacttttgca attgcgacgc
     2581 tgacaaggat gaatggacaa atgatactgg ctttctttcc ttcaaagacc acttgcctgt
     2641 cactcagata gttatcactg ataccgacag atcaaactca gaagccgctt ggagaattgg
     2701 tcccttgcgt tgctatggtg accgacgctt ctggaacgcc gtctcatttt atacagaagc
     2761 ctcttacctc cactttccta ccttccatgc ggaattcagt gccgatattt ccttcttttt
     2821 taaaaccaca gcattatccg gagttttcct agaaaatctt ggcattaaag acttcattcg
     2881 actcgaaata agctctcctt cagagatcac ctttgccatc gatgttggga atggtcctgt
     2941 ggagcttgta gtccagtctc cttctcttct gaatgacaac caatggcact atgtccgggc
     3001 tgagaggaac ctcaaggaga cctccctgca ggtggacaac cttccaagga gcaccaggga
     3061 gacgtcggag gagggccatt ttcgactgca gctgaacagc cagttgtttg taggggggaac
     3121 gtcatccaga cagaaaggct tcctaggatg cattcgctcc ttacacttga atggacagaa
     3181 aatggacctg gaagagaggg caaaggtcac atctggagtc aggccaggct gccccggcca
     3241 ctgcagcagc tacggcagca tctgccacaa cgggggcaag tgtgtggaga agcacaatgg
     3301 ctacctgtgt gattgcacca attcacctta tgaagggccc ttttgcaaaa aagaggtttc
     3361 tgctgttttt gaggctggca cgtcggttac ttacatgttt caagaaccct atcctgtgac
     3421 caagaatata agcctctcat cctcagctat ttacacagat tcagctccat ccaaggaaaa
     3481 cattgcactt agctttgtga caacccaggc acccagtctt ttgctcttta tcaattcttc
     3541 ttctcaggac ttcgtggttg ttctgctctg caagaatgga agcttacagg ttcgctatca
     3601 cctaaacaag gaagaaaccc atgtattcac cattgatgca gataactttg ctaacagaag
     3661 gatgcaccac ttgaagatta accgagaggg aagagagctt accattcaga tggaccagca
     3721 acttcgactc agttataact tctctccgga agtagagttc agggttataa ggtcactcac
     3781 cttgggcaaa gtcacagaga atcttggttt ggattctgaa gttgctaaag caaatgccat
     3841 gggtttgct ggatgcatgt cttccgtcca gtacaaccac atagcaccac tgaaggctgc
     3901 cctgcgccat gccactgtcg cgcctgtgac tgtccatggg accttgacgg aatccagctg
     3961 tggcttcatg gtggactcag atgtgaatgc agtgaccacg gtgcattctt catcagatcc
     4021 ttttgggaag acagatgagc gggaaccact cacaaatgct gttcgaagtg attcggcagt
     4081 catcggaggg gtgatagcag tggtgatatt catcatcttc tgtatcatcg gcatcatgac
     4141 ccggttcctc taccagcaca agcagtcaca tcgtacgagc cagatgaagg agaaggaata
     4201 tccagaaaat ttggacagtt ccttcagaaa tgaaattgac ttgcaaaaca cagtgagcga
     4261 gtgtaaacgg gaatatttca tctgagaaac tgcagggttc ctactactct tttttcttgt
     4321 tgttcaatta tctcctcccc ctcttctctc ctgtcttttg atttggtcat tctctttatt
     4381 ttctgcttgc catgtctttt ctggaacata cttgcatcca ccacagcatc aattcccttg
     4441 atccagccca agagaccagg cagccatggc cactgccttc ctctctgatg aacctatcgg
     4501 gtgaaaacga ccactcaaga gactgacttc gccattcaag acaaggaaga gacacatgtg
     4561 tgcactcctg catgttcagt tctgtacttc cagtttctaa aatgcactgt tcagttttcc
     4621 aaccacttgg tggttcaggc ttgctttgaa cctgagctct taggcacatg acggtcattc
     4681 ctgacatcct ccccagctca agtctattct taccatagaa cccagggcag ggagagaaga
     4741 acctagaggc ctggtttgct ttggtggcat tgtaaaaaga gtaagagagg tttggtttgt
     4801 ggtggtttgc tttctttacc ataagcaatc ccttgcctta actcatcacc ctttttcact
     4861 atgaccctta gaccctgagt attttcaaat atatgattgc tgatagtagt gaccaaaact
     4921 actttgttcc tttcttacca ctctctcctg gggccgacac gttgggacag cacaccatag
     4981 cataaagcta ggggatgcat ggaaatagca gcttgaaact aggaggtaac aagaaagctt
     5041 ctaggaagta gatgttccat atcttcaaaa tgcctcctcc aattttgtaa gaatgctagc
     5101 taggtattcc tgggattatt atactgagat atatatatat acacacacac acacatatgt
     5161 gtatatatgt atatatatat gtgagtatat atacacacac acacacacac acacatatat
     5221 atatatacac acacgcacac atatatgttg ctgcagcata aagaaattga aataaaagtt
     5281 taaa
//
```

Revised: July 5, 2002.

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

```
>AB077881 ACCESSION:AB077881 NID: gi 18181975 dbj AB077881.1  Homo
           sapiens mRNA for caspr5, complete cds
          Length = 4920

 Score = 2567 bits (6581), Expect = 0.0
 Identities = 1303/1307 (99%), Positives = 1303/1307 (99%), Gaps = 3/1307 (0%)
 Frame = +1

Query: 1    MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSPAQLNW 60
            MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSPAQLNW
Sbjct: 1    MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSPAQLNW 180

Query: 61   RVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRNWKQY 120
            RVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRNWKQY
Sbjct: 181  RVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRNWKQY 360

Query: 121  KQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDV 180
            KQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDV
Sbjct: 361  KQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDV 540

Query: 181  ADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKGRLAL 240
            ADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKGRLAL
Sbjct: 541  ADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKGRLAL 720

Query: 241  HLNLGDSKARLSSSLPSATLGSLLDDQHWH-VLIERVGKQVNFTVDKHTQHFRTKGETDA 299
            HLNLGDSKARLSSSLPSATLGSLLDDQHWH VLIERVGKQVNFTVDKHTQHFRTKGETDA
Sbjct: 721  HLNLGDSKARLSSSLPSATLGSLLDDQHWHSVLIERVGKQVNFTVDKHTQHFRTKGETDA 900

Query: 300  LDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNII-LAKRRKHQIYTVGNVTFS 358
            LDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNII LAKRRKHQIYT GNVTFS
Sbjct: 901  LDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNIIDLAKRRKHQIYT-GNVTFS 1077

Query: 359  CSEPQIVPITF-NSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTLLLS 417
            CSEPQIVPITF NSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTLLLS
Sbjct: 1078 CSEPQIVPITFVNSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTLLLS 1257

Query: 418  LEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPDSTW 477
            LEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPDSTW
Sbjct: 1258 LEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPDSTW 1437

Query: 478  VQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNFSDL 537
            VQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNFSDL
Sbjct: 1438 VQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNFSDL 1617

Query: 538  HIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGN 597
            HIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGN
Sbjct: 1618 HIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGN 1797

Query: 598  TAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGG 657
            TAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGG
Sbjct: 1798 TAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGG 1977

Query: 658  SMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQ 717
            SMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQ
Sbjct: 1978 SMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQ 2157
```

```
Query:  718  CECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIG 777
             CECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIG
Sbjct: 2158  CECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIG 2337

Query:  778  PLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIR 837
             PLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIR
Sbjct: 2338  PLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIR 2517

Query:  838  LEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRE 897
             LEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRE
Sbjct: 2518  LEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRE 2697

Query:  898  TSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGH 957
             TSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGH
Sbjct: 2698  TSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGH 2877

Query:  958  CSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVT 1017
             CSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVT
Sbjct: 2878  CSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVT 3057

Query: 1018  KNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYH 1077
             KNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYH
Sbjct: 3058  KNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYH 3237

Query: 1078  LNKEETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLT 1137
             LNKEETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLT
Sbjct: 3238  LNKEETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLT 3417

Query: 1138  LGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSC 1197
             LGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSC
Sbjct: 3418  LGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSC 3597

Query: 1198  GFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMT 1257
             GFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMT
Sbjct: 3598  GFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMT 3777

Query: 1258  RFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI 1304
             RFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI
Sbjct: 3778  RFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI 3918
```

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Boo

Search Nucleotide ▾ for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display default ▾ Save Text Add to Clipboard Get Subsequence

☐ 1: AB077881. Homo sapiens mRNA...[gi:18181975]

Links

```
LOCUS       AB077881                4920 bp    mRNA    linear   PRI 17-JAN-2002
DEFINITION  Homo sapiens mRNA for caspr5, complete cds.
ACCESSION   AB077881
VERSION     AB077881.1  GI:18181975
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Takeuchi,K., Watanabe,N., Kawano,T. and Kawamura,K.
  TITLE     In vitro and in vivo studies on the involvement of neural cell
            adhesion molecules and chondroitin sulfate proteoglycans in
            defining discrete axonal pathways of the rat cerebral cortex
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 4920)
  AUTHORS   Takeuchi,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-JAN-2002) Kosei Takeuchi, Nagoya University, Dept. of
            Biological Sciences; Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602,
            Japan (E-mail:ktakeuch@biol1.bio.nagoya-u.ac.jp,
            Tel:81-52-789-2496, Fax:81-052-789-2968)
FEATURES             Location/Qualifiers
     source          1..4920
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /tissue_type="brain"
     gene            1..4920
                     /gene="caspr5"
     CDS             1..3921
                     /gene="caspr5"
                     /codon_start=1
                     /product="caspr5"
                     /protein_id="BAB83897.1"
                     /db_xref="GI:18181976"
                     /translation="MDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFS
                     SSSDLTGTHSPAQLNWRVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDW
                     VTSYSLMFSDTGRNWKQYKQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWN
                     PSGKIGMRVEVYGCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVL
                     FHGEGQRGDHITLELQKGRLALHLNLGDSKARLSSSLPSATLGSLLDDQHWHSVLIER
                     VGKQVNFTVDKHTQHFRTKGETDALDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLY
                     YNGVNIIDLAKRRKHQIYTGNVTFSCSEPQIVPITFVNSSGSYLLLPGTPQIDGLSVS
                     FQFRTWNKDGLLLSTELSEGSGTLLLSLEGGILRLVIQKMTERVAEILTGSNLNDGLW
                     HSVSINARRNRITLTLDDEAAPPAPDSTWVQIYSGNSYYFGGCPDNLTDSQCLNPIKA
                     FQGCMRLIFIDNQPKDLISVQQGSLGNFSDLHIDLCSIKDRCLPNYCEHGGSCSQSWT
                     TFYCNCSDTSYTGATCHNSIYEQSCEVYRHQGNTAGFFYIDSDGSGPLGPLQVYCNIT
                     EDKIWTSVQHNNTELTRVRGANPEKPYAMALDYGGSMEQLEAVIDGSEHCEQEVAYHC
                     RRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPGVQQCECGLDESCLDIQHFCNCDAD
                     KDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAWRIGPLRCYGDRRFWNAVSFYTE
                     ASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKDFIRLEISSPSEITFAIDVGN
                     GPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRSTRETSEEGHFRLQLNSQL
                     FVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGCPGHCSSYGSICHNGGK
                     CVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVTKNISLSSSAIY
                     TDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQVRYHLNKEETHVF
                     TIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLTLGKVTEN
                     LGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSCGFMVD
                     SDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMTRFL
                     YQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI"
BASE COUNT     1292 a   1229 c   1160 g   1239 t
ORIGIN
        1 atggattctt taccacggct gaccagcgtt ttgactttgc tgttctctgg cttgtggcat
       61 ttaggattaa cagcgacaaa ctacaactgt gatgatccac tagcatccct gctctctcca
      121 atggcttttt ccagttcctc agacctcact ggcactcaca gcccagctca actcaactgg
      181 agagttggaa ctggcggttg gtccccagca gattccaatg ctcaacagtg gctccagatg
      241 gacctgggaa acagagtaga gattacagca gtggccacgc agggaagata cggaagctct
      301 gactgggtga cgagttacag cctgatgttc agtgacacag gacgcaactg gaaacagtac
      361 aaacaagaag acagcatctg gacctttgca ggaaacatga atgctgacag cgtggtgcac
      421 cacaagctat tgcactcagt gagagcccga tttgttcgct ttgtgcccct ggaatggaat
      481 cccagtggga agattggcat gagagtcgag gtctacggat gttcctataa atcagatgtt
      541 gctgactttg atggccgaag ctcacttctg tacaggttca atcagaagtt gatgagtact
      601 ctcaaagatg tgatctccct gaagttcaag agcatgcaag gagatggggt cctgttccat
      661 ggagaaggtc agcgtggaga ccacatcacc ttggaactcc agaaggggag gctcgcccta
      721 cacctcaatt tgggtgacag caaagcgcgg ctcagcagca gcttgccctc tgccaccctg
      781 ggcagcctcc tggatgacca gcactggcac tcggtcctca ttgagcgggt gggcaagcag
      841 gtgaacttca cggtggacaa gcacacacag cacttccgca ccaagggcga gacggatgcc
      901 ttagacattg actatgagct tagttttgga ggaattccag taccaggaaa acctgggacc
      961 tttttaaaga aaaacttcca tggatgcatc gaaaaccttt actacaatgg agtaaacata
     1021 attgacctgg ctaagagacg aaagcatcag atctatactg gcaatgtcac tttttcctgc
     1081 tccgaaccac agattgtgcc catcacattt gtcaactcca gcggcagcta tttgctgctg
     1141 cccggcaccc cccaaattga tgggctctca gtgagtttcc agtttcgaac atggaacaag
     1201 gatggtctgc ttctgtccac agagctgtct gagggctcgg gaaccctgct gctgagcctg
     1261 gagggtggaa tcctgagact cgtgattcag aaaatgacag aacgcgtagc tgaaatcctc
     1321 acaggcagca acttgaatga tggcctgtgg cactcggtta gcatcaacgc caggaggaac
     1381 cgcatcacgc tcactctgga tgatgaagca gcacccccgg ctccagacag cacttgggtg
     1441 cagatttatt ctggaaatag ctactatttt ggagggtgcc ccgacaatct caccgattcc
     1501 caatgtttaa atcccattaa ggctttccaa ggctgcatga ggctcatctt tattgataac
     1561 cagcccaagg acctcatttc agttcagcaa ggttccctgg ggaattttag tgatttacac
     1621 attgatctgt gtagcatcaa agacaggtgt ttgccaaact actgtgaaca tggaggaagc
     1681 tgctcccagt cctggactac cttctattgt aactgcagtg acacaagtta cactggtgcc
     1741 acctgccaca actccatcta cgagcaatcc tgcgaggtgt acaggcacca ggggaataca
     1801 gccggcttct tctacatcga ctcagatggc agcggcccac tgggacctct ccaggtgtac
     1861 tgcaatatca ctgaggacaa gatctggaca tcagtgcagc acaacaatac agagctgacc
     1921 cgagtgcggg gcgctaaccc tgagaagccc tatgccatgg ccttggacta cgggggcagc
     1981 atggaacagc tggaggccgt gatcgacggc tctgagcact gtgagcagga ggtggcctac
     2041 cactgcagga ggtcccgcct gctcaacacg ccggatggaa caccatttac ctggtggatt
     2101 gggcggtcca atgaaaggca cccttactgg ggaggttccc ctcctggggt ccagcagtgt
     2161 gagtgtggcc tagacgagag ctgcctggac attcagcact tttgcaattg cgacgctgac
     2221 aaggatgaat ggacaaatga tactggcttt ctttccttca aagaccactt gcctgtcact
     2281 cagatagtta tcactgatac cgacagatca aactcagaag ccgcttggag aattggtccc
     2341 ttgcgttgct atggtgaccg acgcttctgg aacgccgtct cattttatac agaagcctct
     2401 tacctccact ttcctacctt ccatgcggaa ttcagtgccg atatttcctt ctttttaaa
     2461 accacagcat tatccggagt tttcctagaa aatcttggca ttaaagactt cattcgactc
     2521 gaaataagct ctccttcaga gatcaccttt gccatcgatg ttgggaatgg tcctgtggag
     2581 cttgtagtcc agtctccttc tcttctgaat gacaaccaat ggcactatgt ccgggctgag
     2641 aggaacctca aggagacctc cctgcaggtg gacaaccttc caaggagcac cagggagacg
     2701 tcggaggagg gccattttcg actgcagctg aacagccagt tgtttgtagg gggaacgtca
     2761 tccagacaga aaggcttcct aggatgcatt cgctccttac acttgaatgg acagaaaatg
     2821 gacctggaag agagggcaaa ggtcacatct ggagtcaggc caggctgccc cggccactgc
     2881 agcagctacg gcagcatctg ccacaacggg ggcaagtgtg tggagaagca caatggctac
     2941 ctgtgtgatt gcaccaattc accttatgaa gggccctttt gcaaaaaaga ggtttctgct
     3001 gtttttgagg ctggcacgtc ggttacttac atgtttcaag aaccctatcc tgtgaccaag
     3061 aatataagcc tctcatcctc agctatttac acagattcag ctccatccaa ggaaaacatt
     3121 gcacttagct ttgtgacaac ccaggcaccc agtcttttgc tctttatcaa ttcttcttct
     3181 caggacttcg tggttgttct gctctgcaag aatggaagct tacaggttcg ctatcaccta
     3241 aacaaggaag aaacccatgt attcaccatt gatgcagata actttgctaa cagaaggatg
     3301 caccacttga agattaaccg agagggaaga gagcttacca ttcagatgga ccagcaactt
     3361 cgactcagtt ataacttctc tccggaagta gagttcaggg ttataaggtc actcaccttg
     3421 ggcaaagtca cagagaatct tggtttggat tctgaagttg ctaaagcaaa tgccatgggt
     3481 tttgctggat gcatgtcttc cgtccagtac aaccacatag caccactgaa ggctgccctg
     3541 cgccatgcca ctgtcgcgcc tgtgactgtc catgggacct tgacggaatc cagctgtggc
     3601 ttcatggtgg actcagatgt gaatgcagtg accacggtgc attcttcatc agatcctttt
     3661 gggaagacag atgagcggga accactcaca aatgctgttc gaagtgattc ggcagtcatc
     3721 ggaggggtga tagcagtggt gatattcatc atcttctgta tcatcggcat catgacccgg
     3781 ttcctctacc agcacaagca gtcacatcgt acgagccaga tgaaggagaa ggaatatcca
     3841 gaaaatttgg acagttcctt cagaaatgaa attgacttgc aaaacacagt gagcgagtgt
     3901 aaacgggaat atttcatctg agaaactgca gggttcctac tactcttttt tcttgttgtt
     3961 caattatctc ctccccctct tctctcctgt cttttgattt ggtcattctc tttattttct
     4021 gcttgccatg tcttttctgg aacatacttg catccaccac agcatcaatt cccttgatcc
     4081 agcccaagag accaggcagc catggccact gccttcctct ctgatgaacc tatcgggtga
     4141 aaacgaccac tcaagagact gacttcgcca ttcaagacaa ggaagagaca catgtgtgca
     4201 ctcctgcatg ttcagttctg tacttccagt ttctaaaatg cactgttcag ttttccaacc
     4261 acttggtggt tcaggcttgc tttgaacctg agctcttagg cacatgacgg tcattcctga
     4321 catcctcccc agctcaagtc tattcttacc atagaaccca gggcagggag agaagaacct
     4381 agaggcctgg tttgctttgg tggcattgta aaaagagtaa gagaggtttg gtttgtggtg
     4441 gtttgctttc tttaccataa gcaatccctt gccttaactc atcacccttt ttcactatga
     4501 cccttagacc ctgagtattt tcaaatatat gattgctgat agtagtgacc aaaactactt
     4561 tgttcctttc ttaccactct ctcctggggc cgacacgttg ggacagcaca ccatagcata
     4621 aagctagggg atgcatggaa atagcagctt gaaactagga ggtaacaaga aagcttctag
     4681 gaagtagatg ttccatatct tcaaaatgcc tcctccaatt ttgtaagaat gctagctagg
     4741 tattcctggg attattatac tgagatatat atatatacac acacacacac atatgtgtat
     4801 atatgtatat atatatgtga gtatatatac acacacacac acacacacac atatatatat
     4861 atacacacac gcacacatat atgttgctgc agcataaaga aattgaaata aaagtttaaa
//
```

Revised: July 5, 2002.

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Dec 13 2002 14:41:17

NCBI PubMed National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Bc

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

Display Abstract Show: 20 Sort Send to File

□ 1: Neuron 1999 Dec;24(4):1037-47 Related Articles, Links

FULL TEXT @ CELL PRESS

Caspr2, a new member of the neurexin superfamily, is localized at the juxtaparanodes of myelinated axons and associates with K+ channels.

Poliak S, Gollan L, Martinez R, Custer A, Einheber S, Salzer JL, Trimmer JS, Shrager P, Peles E.

Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot, Israel.

Rapid conduction in myelinated axons depends on the generation of specialized subcellular domains to which different sets of ion channels are localized. Here, we describe the identification of Caspr2, a mammalian homolog of Drosophila Neurexin IV (Nrx-IV), and show that this neurexin-like protein and the closely related molecule Caspr/Paranodin demarcate distinct subdomains in myelinated axons. While contactin-associated protein (Caspr) is present at the paranodal junctions, Caspr2 is precisely colocalized with Shaker-like K+ channels in the juxtaparanodal region. We further show that Caspr2 specifically associates with Kv1.1, Kv1.2, and their Kvbeta2 subunit. This association involves the C-terminal sequence of Caspr2, which contains a putative PDZ binding site. These results suggest a role for Caspr family members in the local differentiation of the axon into distinct functional subdomains.

PMID: 10624965 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to File

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Freedom of Information Act | Disclaimer

i686-pc-linux-gnu Dec 13 2002 14:22:59

NCBI PubMed National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Bc

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Display Abstract Show: 20 Sort Send to File

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

□ 1: Mol Cell Neurosci 2002 Jun;20(2):283-97 Related Articles, Links

ELSEVIER SCIENCE
FULL-TEXT ARTICLE

Caspr3 and caspr4, two novel members of the caspr family are expressed in the nervous system and interact with PDZ domains.

Spiegel I, Salomon D, Erne B, Schaeren-Wiemers N, Peles E.

Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot 76100, Israel.

The NCP family of cell-recognition molecules represents a distinct subgroup of the neurexins that includes Caspr and Caspr2, as well as Drosophila Neurexin-IV and axotactin. Here, we report the identification of Caspr3 and Caspr4, two new NCPs expressed in nervous system. Caspr3 was detected along axons in the corpus callosum, spinal cord, basket cells in the cerebellum and in peripheral nerves, as well as in oligodendrocytes. In contrast, expression of Caspr4 was more restricted to specific neuronal subpopulations in the olfactory bulb, hippocampus, deep cerebellar nuclei, and the substantia nigra. Similar to the neurexins, the cytoplasmic tails of Caspr3 and Caspr4 interacted differentially with PDZ domain-containing proteins of the CASK/Lin2-Veli/Lin7-Mint1/Lin10 complex. The structural organization and distinct cellular distribution of Caspr3 and Caspr4 suggest a potential role of these proteins in cell recognition within the nervous system. (c) 2002 Elsevier Science (USA).

PMID: 12093160 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to File

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Freedom of Information Act | Disclaimer

i686-pc-linux-gnu Dec 13 2002 14:22:59

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAa9aWNs: 1331 aa
 >gi|6694278|gb|AAF25199.1|AF193613_1 cell recognition molecule Caspr2 [Homo sapiens
 vs  /tmp/fastaDAAb9aWNs library
searching /tmp/fastaDAAb9aWNs library

   1154 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                          opt
gi|16306509|ref|NP_387504.1| cell recognition mol  (1154) 1595

>>gi|16306509|ref|NP_387504.1| cell recognition molecule  (1154 aa)
 initn: 3285 init1: 1327 opt: 1595
Smith-Waterman score: 3418;  41.992% identity in 1305 aa overlap (12-1305:2-1153)

               10        20              30        40        50
gi|669 MQAAPRAGCGAALLLWIVSSCL----CRAWT--APSTSQKCDEPLVSGLPHVAFSSSSSI
                  : . : : . :     ..:.  . ..   :: ::.:.::. .::::: .
gi|163           MASVAWAVLKVLLLLPTQTWSPVGAGNPPDCDAPLASALPRSSFSSSSEL
                         10        20        30        40        50

           60        70        80        90       100       110
gi|669 SGSYSPGYAKINKRGGAGGWSPSDSDHYQWLQVDFGNRKQISAIATQGRYSSSDWVTQYR
       :.:..::....:.: :::::.:  :..:::::.:.:.: ...:.:::: :.::::::.:
gi|163 SSSHGPGFSRLNRRDGAGGWTPLVSNKYQWLQIDLGERMEVTAVATQGGYGSSDWVTSYL
               60        70        80        90       100       110

          120       130       140       150       160       170
gi|669 MLYSDTGRNWKPYHQDGNIWAFPGNINSDGVVRHELQHPIIARYVRIVPLDWNGEGRIGL
       ...:: ::::: :... .::.:::: :.:.::...:: :. ::..:..:: :: .::::.
gi|163 LMFSDGGRNWKQYRREESIWGFPGNTNADSVVHYRLQPPFEARFLRFLPLAWNPRGRIGM
              120       130       140       150       160       170

          180       190       200       210       220       230
gi|669 RIEVYGCSYWADVINFDGHVVLPYRFRNKKMKTLKDVIALNFKTSESEGVILHGEGQQGD
       :::::::.: ..:. :::. .: ::. .: .: ..:::.:.::. .:.:..:: :::.:.
gi|163 RIEVYGCAYKSEVVYFDGQSALLYRLDKKPLKPIRDVISLKFKAMQSNGILLHREGQHGN
              180       190       200       210       220       230

          240       250       260       270       280       290
gi|669 YITLELKKAKLVLSLNLGSNQLGPIYGHTSVMTGSLLDDHHWHSVVIERQGRSINLTLDR
       .::::: :.:::. :: :. .:    . ...  ::::::.:::::.::    ..:.:.:.
gi|163 HITLELIKGKLVFFLNSGNAKLPSTIAPVTLTLGSLLDDQHWHSVLIELLDTQVNFTVDK
              240       250       260       270       280       290

          300       310       320       330       340       350
gi|669 SMQHFRTNGEFDYLDLDYEITFGGIPFSGKPSSSSRKNFKGCMESINYNGVNITDLARRK
         .::...:. .::::..::.:::::  :.  .  ::.:.::.:.. ::::..:.::...
gi|163 HTHHFQAKGDSSYLDLNFEISFGGIPTPGRSRAFRRKSFHGCLENLYYNGVDVTELAKKH
              300       310       320       330       340       350

          360       370        380       390       400       410
gi|669 KLEPSNVGNLSFSCVEPYTVPV-FFNATSYLEVPGRLNQDLFSVSFQFRTWNPNGLLVFS
```

```
       : .   .::.:::: .: :::: :... ::: .::  ..:  ::.:::::::  : :.:.
gi|163 KPQILMMGNVSFSCPQPQTVPVTFLSSRSYLALPGNSGEDKVSVTFQFRTWNRAGHLLFG
              360       370       380       390       400       410

           420       430       440       450       460       470
gi|669 HFADNLGNVEIDLTESKVGVHINITQTKMSQIDISSGSGLNDGQWHEVRFLAKENFAILT
       .. . :.  . : ..:.  .... :  .:  ....:.:::::::: : : :: .   ..
gi|163 ELRRGSGSFVLFLKDGKL--KLSLFQPGQSPRNVTAGAGLNDGQWHSVSFSAKWSHMNVV
              420         430       440       450       460

           480       490       500       510       520       530
gi|669 IDGDEASAVRTNSPLQVKTGEKYFFGGFLNQMNNSSHSVLQPSFQGCMQLIQVDDQLVNL
       .: :  .::.    . . .:. :.::       ... .:.:
gi|163 VDDD--TAVQPLVAVLIDSGDTYYFG-------DAAWTVVQ-------------------
      470         480       490              500

           540       550       560       570       580       590
gi|669 YEVAQRKPGSFANVSIDMCAIIDRCVPNHCEHGGKCSQTWDSFKCTCDETGYSGATCHNS
                                      :::             : .   ::
gi|163 -------------------------------HGGP------------DAVTLRGA-----
                                                         510

           600       610       620       630       640       650
gi|669 IYEPSCEAYKHLGQTSNYYWIDPDGSGPLGPLKVYCNMTEDKVWTIVSHDLQMQTPVVGY
                             :.:
gi|163 ----------------------PSG-----------------------------------


           660       670       680       690       700       710
gi|669 NPEKYSVTQLVYSASMDQISAITDSAEYCEQYVSYFCKMSRLLNTPDGSPYTWWVGKANE
       .:.  :.....:.:.  :. . .. :: ::: ..  :  .:  .. ::.: .::::..::
gi|163 HPR--SAVSFAYAAGAGQLRSAVNLAERCEQRLALRCGTARRPDSRDGTPLSWWVGRTNE
           520       530       540       550       560       570

           720       730       740       750       760       770
gi|669 KHYYWGGSGPGIQKCACGIERNCTDPKYYCNCDADYKQWRKDAGFLSYKDHLPVSQVVVG
        : ::::: :  :::.::.: :: : .:::::::  ..: .:.  :: :.::::.:.:.
gi|163 THTYWGGSLPDAQKCTCGLEGNCIDSQYYCNCDAGRNEWTSDTIVLSQKEHLPVTQIVMT
           580       590       600       610       620       630

           780       790       800       810       820       830
gi|669 DTDRQGSEAKLSVGPLRCQGDRNYWNAASFPNPSSYLHFSTFQGETSADISFYFKTLTPW
       :: .  :::  ..::: :.::...::.::: . .::::: .:.:: .::. :.::: .
gi|163 DTGQPHSEADYTLGPLLCRGDQSFWNSASFNTETSYLHFPAFHGELTADVCFFFKTTVSS
           640       650       660       670       680       690

           840       850       860       870       880       890
gi|669 GVFLENMGKEDFIKLELKSATEVSFSFDVGNGPVEIVVRSPTPLNDDQWHRVTAERNVKQ
       :::.::.:  :::..::.. :::.::::::::: :..:.::::.::.:::.: ::::::
gi|163 GVFMENLGITDFIRIELRAPTEVTFSFDVGNGPCEVTVQSPTPFNDNQWHHVRAERNVKG
           700       710       720       730       740       750

           900       910       920        930       940       950
gi|669 ASLQVDRLPQQIRKAPTEGHTRLELYSQLFVGG-AGGQQGFLGCIRSLRMNGVTLDLEER
       ::::::.:::... ::..::.::.: ::::.:: :  :.:::::::::..:::.::::::
gi|163 ASLQVDQLPQKMQPAPADGHVRLQLNSQLFIGGTATRQRGFLGCIRSLQLNGVALDLEER
           760       770       780       790       800       810

            960       970       980       990      1000      1010
gi|669 AKVTSGFISGCSGHCTSYGTNCENGGKCLERYHGYSCDCSNTAYDGTFCNKDVGAFFEEG
```

```
       : :: :   ::.:::..::  :.:::.: :. .: .:::. .:::: ::.....:.:  :
gi|163 ATVTPGVEPGCAGHCSTYGHLCRNGGRCREKRRGVTCDCAFSAYDGPFCSNEISAYFATG
           820       830       840       850       860       870

           1020      1030      1040        1050      1060      1070
gi|669 MWLRYNFQAPATNARDSSSRVDNAPDQQNSHPD--LAQEEIRFSFSTTKAPCILLYISSF
         . :.::   : ...::: :..       : :  :..: : .:: ::..: .:::.:::
gi|163 SSMTYHFQEHYTLSENSSSLVSSL------HRDVTLTREMITLSFRTTRTPSLLLYVSSF
           880       890             900       910       920

             1080      1090      1100      1110      1120      1130
gi|669 TTDFLAVLVKPTGSLQIRYNLGGTREPYNIDVDHRNMANGQPHSVNITRHEKTIFLKLDH
         ..:.:..  .:::::::.:   ..:  .  : .:::.:: :.:.:.:   ........
gi|163 YEEYLSVILANNGSLQIRYKLDRHQNPDAFTFDFKNMADGQLHQVKINREEAVVMVEVNQ
       930       940       950       960       970       980

             1140      1150      1160      1170      1180      1190
gi|669 YPSVSYHLPSSSDTLFNSPKSLFLGKVIETGKIDQEIHKYNTPGFTGCLSRVQFNQIAPL
         :..  ..  :: : ::. :::.::::.:..  : .  ..  : ::::::: :.:.. :::
gi|163 --STKKQVILSSGTEFNAVKSLILGKVLEAAGADPDTRRAATSGFTGCLSAVRFGRAAPL
         990      1000      1010      1020      1030      1040

             1200      1210       1220      1230      1240
gi|669 KAALRQTNASAHVHIQGELVE-SNCGASPLTLSPMSSATDPWHLDHLDSASADFPYNPGQ
       :::::  .. : .: ..:...  . :.:.  . ::    .         :. ::      .
gi|163 KAALRPSGPS-RVTVRGHVAPMARCAAGAASGSPARELAPRLAGGAGRSGPAD------E
        1050       1060      1070      1080      1090

    1250      1260      1270      1280      1290      1300
gi|669 GQAIRNGVNRNSAIIGGVIAVVIFTILCTLVFLIRYMFRHKGTYHTNEAKGAESAESADA
       :. . :.  :.::.::::::::::: .::  ..  :: .....   .  ::.:  ... :
gi|163 GEPLVNADRRDSAVIGGVIAVVIFILLCITAIAIR-IYQQRKLRKENESKVSKKEEC
     1100      1110      1120      1130       1140      1150

    1310      1320      1330
gi|669 AIMNNDPNFTETIDESKKEWLI


1331 residues in 1 query   sequences
1154 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:41:43 2002 done: Mon Dec 16 15:41:44 2002
 Scan time:  0.050 Display time:  2.167

Function used was FASTA
```

E

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaOAA78aisO: 1331 aa
 >gi|6694278|gb|AAF25199.1|AF193613_1 cell recognition molecule Caspr2 [Homo sapiens
 vs  /tmp/fastaPAA88aisO library
searching /tmp/fastaPAA88aisO library

   1311 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                      opt
gi|18496979|ref|NP_207837.1| cell recognition pro  (1311) 4248

>>gi|18496979|ref|NP_207837.1| cell recognition protein   (1311 aa)
 initn: 3152 init1: 1339 opt: 4248
Smith-Waterman score: 4406;  48.304% identity in 1327 aa overlap (13-1330:7-1310)

               10        20        30             40        50
gi|669 MQAAPRAGCGAALLLWIVSSCLCRAWTAPSTS-----QKCDEPLVSGLPHVAFSSSSSIS
                   :..  ..:    : : :...     . ::.::::.::...::::: .:
gi|184       MLLFYLLVVLSIDSTKASALTNPNVALFLLADDCDDPLVSALPQASFSSSSELS
                      10        20        30        40        50

          60        70        80        90       100       110
gi|669 GSYSPGYAKINKRGGAGGWSPSDSDHYQWLQVDFGNRKQISAIATQGRYSSSDWVTQYRM
       .:..::.:..:.: :::::::  :..:::::.:.:.: ...:.:::: :.::.:::.: .
gi|184 SSHGPGFARLNRRDGAGGWSPLVSNKYQWLQIDLGERMEVTAVATQGGYGSSNWVTSYLL
           60        70        80        90       100       110

         120       130       140       150       160       170
gi|669 LYSDTGRNWKPYHQDGNIWAFPGNINSDGVVRHELQHPIIARYVRIVPLDWNGEGRIGLR
       ..::.: ::: :.:. .::.: :: :.:.:: ..::  : ::..:..::.:: .::::.:
gi|184 MFSDSGWNWKQYRQEDSIWGFSGNANADSVVYYRLQPSIKARFLRFIPLEWNPKGRIGMR
          120       130       140       150       160       170

         180       190       200       210       220       230
gi|669 IEVYGCSYWADVINFDGHVVLPYRFRNKKMKTLKDVIALNFKTSESEGVILHGEGQQGDY
       :::.::.: ..:...::.  : ::: .:... .::.:.:.::: .:.:..:: :: .::.
gi|184 IEVFGCAYRSEVVDLDGKSSLLYRFDQKSLSPIKDIISLKFKTMQSDGILLHREGPNGDH
          180       190       200       210       220       230

         240       250       260       270       280       290
gi|669 ITLELKKAKLVLSLNLGSNQLGPIYGHTSVMTGSLLDDHHWHSVVIERQGRSINLTLDRS
       :::.:..:.: : .: :  .:      ...  ::::::.:::::.:.: :...:.:.:.
gi|184 ITLQLRRARLFLLINSGEAKLPSTSTLVNLTLGSLLDDQHWHSVLIQRLGKQVNFTVDEH
          240       250       260       270       280       290

         300       310       320       330       340       350
gi|669 MQHFRTNGEFDYLDLDYEITFGGIPFSGKPSSSSRKNFKGCMESINYNGVNITDLARRKK
        .::.. :::. ..:::::.:::::  ::  :  ..::.::.:.. ::::.: :::...:
gi|184 RHHFHARGEFNLMNLDYEISFGGIPAPGKSVSFPHRNFHGCLENLYYNGVDIIDLAKQQK
          300       310       320       330       340       350

         360       370        380       390       400       410
gi|669 LEPSNVGNLSFSCVEPYTVPV-FFNATSYLEVPGRLNQDLFSVSFQFRTWNPNGLLVFSH
```

```
        .   .::.:::: .: ..:: :... ::: .:   ...  :..:::::::  :::.::.
gi|184 PQIIAMGNVSFSCSQPQSMPVTFLSSRSYLALPDFSGEEEVSATFQFRTWNKAGLLLFSE
          360       370       380       390       400       410

          420       430       440       450       460       470
gi|669 FADNLGNVEIDLTESKVGVHINITQTKMSQIDISSGSGLNDGQWHEVRFLAKENFAILTI
       .    :.. . :...:.  . :. :      ::..:  ::::::: : . ::.:   ...
gi|184 LQLISGGILLFLSDGKL--KSNLYQPGKLPSDITAGVELNDGQWHSVSLSAKKNHLSVAV
          420       430         440       450       460       470

          480       490       500       510       520       530
gi|669 DGDEASAVRTNSPLQVKTGEKYFFGGFLNQMNNSSHSVLQPSFQGCMQLIQVDDQLVNLY
       ::. :::.   .: :. .:  :.:::  ..  .:. .    .:::::.::... ..:.:
gi|184 DGQMASAAPLLGPEQIYSGGTYYFGGCPDKSFGSKCKSPLGGFQGCMRLISISGKVVDLI
            480       490       500       510       520       530

          540       550       560       570       580       590
gi|669 EVAQRKPGSFANVSIDMCAIIDRCVPNHCEHGGKCSQTWDSFKCTCDETGYSGATCHNSI
        : : . :.:....:: :.: :::.::.:::::.:::.:..:.:.: .::: ::::::::
gi|184 SVQQGSLGNFSDLQIDSCGISDRCLPNYCEHGGECSQSWSTFHCNCTNTGYRGATCHNSI
            540       550       560       570       580       590

          600       610       620       630       640       650
gi|669 YEPSCEAYKHLGQTSNYYWIDPDGSGPLGPLKVYCNMTEDKVWTIVSHDLQMQTPVVGYN
       :: ::::::: :.::..:.:: :::::: :. .::::::  .:::..:. .  : : . :
gi|184 YEQSCEAYKHRGNTSGFYYIDSDGSGPLEPFLLYCNMTET-AWTIIQHNGSDLTRVRNTN
            600       610       620       630        640       650

          660       670       680       690       700       710
gi|669 PEKYSVTQLVYSASMDQISAITDSAEYCEQYVSYFCKMSRLLNTPDGSPYTWWVGKANEK
       ::.  .  . : :::.:..:  . ::.:::  .:.:: :::.:  ::.: .::::..::
gi|184 PENPYAGFFEYVASMEQLQATINRAEHCEQEFTYYCKKSRLVNKQDGTPLSWWVGRTNET
             660       670       680       690       700       710

          720       730       740       750       760       770
gi|669 HYYWGGSGPGIQKCACGIERNCTDPKYYCNCDADYKQWRKDAGFLSYKDHLPVSQVVVGD
       . :::::.: .:::.::.: :: : .:::::::: ..: .:.:.:.::.::::...:. :
gi|184 QTYWGGSSPDLQKCTCGLEGNCIDSQYYCNCDADRNEWTNDTGLLAYKEHLPVTKIVITD
             720       730       740       750       760       770

          780       790       800       810       820       830
gi|669 TDRQGSEAKLSVGPLRCQGDRNYWNAASFPNPSSYLHFSTFQGETSADISFYFKTLTPWG
       : :  :::  ..::: :.:::..::.::: . .::::: ::.:: :::.::.::: .  :
gi|184 TGRLHSEAAYKLGPLLCRGDRSFWNSASFDTEASYLHFPTFHGELSADVSFFFKTTASSG
             780       790       800       810       820       830

          840       850       860       870       880       890
gi|669 VFLENMGKEDFIKLELKSATEVSFSFDVGNGPVEIVVRSPTPLNDDQWHRVTAERNVKQA
       :::::.:  :::..::.: : :.::::::::: :: :.::: .::.:::.: .:::.:.:
gi|184 VFLENLGIADFIRIELRSPTVVTFSFDVGNGPFEISVQSPTHFNDNQWHHVRVERNMKEA
             840       850       860       870       880       890

          900       910       920        930       940       950
gi|669 SLQVDRLPQQIRKAPTEGHTRLELYSQLFVGG-AGGQQGFLGCIRSLRMNGVTLDLEERA
       :::::.:  . . ::..::. :.: ::::::: :  :.:::::::::..::.::::::::
gi|184 SLQVDQLTPKTQPAPADGHVLLQLNSQLFVGGTATRQRGFLGCIRSLQLNGMTLDLEERA
             900       910       920       930       940       950

           960       970       980       990       1000     1010
gi|669 KVTSGFISGCSGHCTSYGTNCENGGKCLERYHGYSCDCSNTAYDGTFCNKDVGAFFEEGM
```

```
       .::     :: :::.:::  :.::::: ::  :. :::. .:: : ::.....:.:  :
gi|184 QVTPEVQPGCRGHCSSYGKLCRNGGKCRERPIGFFCDCTFSAYTGPFCSNEISAYFGSGS
             960       970       980       990      1000      1010

          1020      1030      1040        1050      1060      1070
gi|669 WLRYNFQAPATNARDSSSRVDNAPDQQNSHPD--LAQEEIRFSFSTTKAPCILLYISSFT
        . ::::     ...:::..       . : :  :..: :.::: ::..: .::..:::
gi|184 SVIYNFQENYLLSKNSSSHA------ASFHGDMKLSREMIKFSFRTTRTPSLLLFVSSFY
            1020      1030            1040      1050      1060

            1080      1090      1100      1110      1120      1130
gi|669 TDFLAVLVKPTGSLQIRYNLGGTREPYNIDVDHRNMANGQPHSVNITRHEKTIFLKLDHY
        ..:.:..  .:::::::.:.  .::  .. : .:::.:: : . :.:.: ..:...:
gi|184 KEYLSVIIAKNGSLQIRYKLNKYQEPDVVNFDFKNMADGQLHHIMINREEGVVFIEIDDN
        1070      1080      1090      1100      1110      1120

            1140      1150      1160      1170      1180      1190
gi|669 PSVSYHLPSSSDTLFNSPKSLFLGKVIETGKIDQEIHKYNTPGFTGCLSRVQFNQIAPLK
          . ::  :: : :.. ::: ::...: . .:::    .. ::::::: ::....::::
gi|184 RRRQVHL--SSGTEFSAVKSLVLGRILEHSDVDQETALAGAQGFTGCLSAVQLSHVAPLK
        1130        1140      1150      1160      1170      1180

            1200      1210      1220      1230      1240      1250
gi|669 AALRQTNASAHVHIQGELVESNCGASPLTLSPMSSATDPWHLDHLDSASADFPYNPGQGQ
       :::. .. .  : . :...::.: :.: : .     :  .  ::  :.. :   .:
gi|184 AALHPSHPDP-VTVTGHVTESSCMAQPGTDATSRERTHSFA-DH--SGTID-DREP----
          1190       1200      1210      1220         1230

            1260      1270      1280      1290      1300      1310
gi|669 AIRNGVNRNSAIIGGVIAVVIFTILCTLVFLIRYMFRHKGTYHTNEAKGAESAESADAAI
        . :... .::.:::.:::::: .::  .. .: ....:  :. .::: .:...::.: .
gi|184 -LANAIKSDSAVIGGLIAVVIFILLCITAIAVR-IYQQKRLYKRSEAKRSENVDSAEA-V
          1240      1250      1260       1270      1280      1290

            1320      1330
gi|669 MNNDPNFTETIDESKKEWLI
       .... :. ....:..::...
gi|184 LKSELNIQNAVNENQKEYFF
            1300      1310


1331 residues in 1 query   sequences
1311 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:45:32 2002 done: Mon Dec 16 15:45:34 2002
 Scan time:  0.050 Display time:  2.484

Function used was FASTA
```

F

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaKAA38aisO: 1311 aa
 >gi|18496979|ref|NP_207837.1| cell recognition protein CASPR4, isoform 1; contactin
 vs  /tmp/fastaLAA48aisO library
searching /tmp/fastaLAA48aisO library

   1154 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                      opt
gi|16306509|ref|NP_387504.1| cell recognition mol  (1154) 3080

>>gi|16306509|ref|NP_387504.1| cell recognition molecule  (1154 aa)
 initn: 5264 init1: 2744 opt: 3080
Smith-Waterman score: 5135;  62.580% identity in 1256 aa overlap (33-1283:30-1150)

             10        20        30        40        50        60
gi|184 LFYLLVVLSIDSTKASALTNPNVALFLLADDCDDPLVSALPQASFSSSSELSSSHGPGFA
                                     ::: ::.::::..::::::::::::::::.
gi|163  MASVAWAVLKVLLLLPTQTWSPVGAGNPPDCDAPLASALPRSSFSSSSELSSSHGPGFS
                10        20        30        40        50

             70        80        90       100       110       120
gi|184 RLNRRDGAGGWSPLVSNKYQWLQIDLGERMEVTAVATQGGYGSSNWVTSYLLMFSDSGWN
       :::::::::::.::::::::::::::::::::::::::::::::.:::::::::::.: :
gi|163 RLNRRDGAGGWTPLVSNKYQWLQIDLGERMEVTAVATQGGYGSSDWVTSYLLMFSDGGRN
      60        70        80        90       100       110

            130       140       150       160       170       180
gi|184 WKQYRQEDSIWGFSGNANADSVVYYRLQPSIKARFLRFIPLEWNPKGRIGMRIEVFGCAY
       :::::.:.::::: ::.::::::.::::: ..::::::.:: :::.:::::::::.::::
gi|163 WKQYRREESIWGFPGNTNADSVVHYRLQPPFEARFLRFLPLAWNPRGRIGMRIEVYGCAY
     120       130       140       150       160       170

            190       200       210       220       230       240
gi|184 RSEVVDLDGKSSLLYRFDQKSLSPIKDIISLKFKTMQSDGILLHREGPNGDHITLQLRRA
       .:::: .::.:.::::.:.: :.::.:.::::::.:::.:::::::: .:.::::.: ..
gi|163 KSEVVYFDGQSALLYRLDKKPLKPIRDVISLKFKAMQSNGILLHREGQHGNHITLELIKG
     180       190       200       210       220       230

            250       260       270       280       290       300
gi|184 RLFLLINSGEAKLPSTSTLVNLTLGSLLDDQHWHSVLIQRLGKQVNFTVDEHRHHFHARG
       .: ...:::.:::::: . :.:::::::::::::::::. :  :::::::.: :::.:.:
gi|163 KLVFFLNSGNAKLPSTIAPVTLTLGSLLDDQHWHSVLIELLDTQVNFTVDKHTHHFQAKG
     240       250       260       270       280       290

            310       320       330       340       350       360
gi|184 EFNLMNLDYEISFGGIPAPGKSVSFPHRNFHGCLENLYYNGVDIIDLAKQQKPQIIAMGN
       . . ..:..::::::::.::.: .: ...::::::::::::::. .:::..::::. :::
gi|163 DSSYLDLNFEISFGGIPTPGRSRAFRRKSFHGCLENLYYNGVDVTELAKKHKPQILMMGN
     300       310       320       330       340       350

            370       380       390       400       410       420
gi|184 VSFSCSQPQSMPVTFLSSRSYLALPDFSGEEEVSATFQFRTWNKAGLLLFSELQLISGGI
```

```
       ::::: :::..:::::::::::::::  :::..::.::::::::.:: :::.::.  ::..
gi|163 VSFSCPQPQTVPVTFLSSRSYLALPGNSGEDKVSVTFQFRTWNRAGHLLFGELRRGSGSF
     360       370       380       390       400       410

            430       440       450       460       470       480
gi|184 LLFLSDGKLKSNLYQPGKLPSDITAGVELNDGQWHSVSLSAKKNHLSVAVDGQMASAAPL
       .:::.::::: .:.:::. : ..:::. ::::::::::.::: .:..:.:: . : . ::
gi|163 VLFLKDGKLKLSLFQPGQSPRNVTAGAGLNDGQWHSVSFSAKWSHMNVVVDDDTA-VQPL
     420       430       440       450       460       470

            490       500       510       520       530       540
gi|184 LGPEQIYSGGTYYFGGCPDKSFGSKCKSPLGGFQGCMRLISISGKVVDLISVQQGSLGNF
       ..   : :: :::::
gi|163 VAV-LIDSGDTYYFG---------------------------------------------
      480       490

            550       560       570       580       590       600
gi|184 SDLQIDSCGISDRCLPNYCEHGGECSQSWSTFHCNCTNTGYRGATCHNSIYEQSCEAYKH

gi|163 ------------------------------------------------------------


            610       620       630       640       650       660
gi|184 RGNTSGFYYIDSDGSGPLEPFLLYCNMTETAWTIIQHNGSDLTRVRNTNPENPYAGF-FE
                                   ..:::..::.: : . .:..   .: ..  : 
gi|163 ----------------------------DAAWTVVQHGGPDAVTLRGAPSGHPRSAVSFA
                                        500       510       520

             670       680       690       700       710       720
gi|184 YVASMEQLQATINRAEHCEQEFTYYCKKSRLVNKQDGTPLSWWVGRTNETQTYWGGSSPD
       :.:.  ::....: ::.:::...  :  .:  ...:::::::::::::::.:::::: ::
gi|163 YAAGAGQLRSAVNLAERCEQRLALRCGTARRPDSRDGTPLSWWVGRTNETHTYWGGSLPD
          530       540       550       560       570       580

             730       740       750       760       770       780
gi|184 LQKCTCGLEGNCIDSQYYCNCDADRNEWTNDTGLLAYKEHLPVTKIVITDTGRLHSEAAY
        :::::::::::::::::::::: :::::.:: .:. :::::::.::.::::. :::: :
gi|163 AQKCTCGLEGNCIDSQYYCNCDAGRNEWTSDTIVLSQKEHLPVTQIVMTDTGQPHSEADY
          590       600       610       620       630       640

             790       800       810       820       830       840
gi|184 KLGPLLCRGDRSFWNSASFDTEASYLHFPTFHGELSADVSFFFKTTASSGVFLENLGIAD
        :::::::::.::::::::.::.::::::.:::::.::: ::::::.:::::.:::::.:
gi|163 TLGPLLCRGDQSFWNSASFNTETSYLHFPAFHGELTADVCFFFKTTVSSGVFMENLGITD
          650       660       670       680       690       700

             850       860       870       880       890       900
gi|184 FIRIELRSPTVVTFSFDVGNGPFEISVQSPTHFNDNQWHHVRVERNMKEASLQVDQLTPK
       :::::::.:: ::::::::::: :..::::: ::::::::::.:::.: ::::::::  :
gi|163 FIRIELRAPTEVTFSFDVGNGPCEVTVQSPTPFNDNQWHHVRAERNVKGASLQVDQLPQK
          710       720       730       740       750       760

             910       920       930       940       950       960
gi|184 TQPAPADGHVLLQLNSQLFVGGTATRQRGFLGCIRSLQLNGMTLDLEERAQVTPEVQPGC
        ::::::::: ::::::::.:::::::::::::::::::::..::::::: ::: :.:::
gi|163 MQPAPADGHVRLQLNSQLFIGGTATRQRGFLGCIRSLQLNGVALDLEERATVTPGVEPGC
          770       780       790       800       810       820

             970       980       990      1000      1010      1020
gi|184 RGHCSSYGKLCRNGGKCRERPIGFFCDCTFSAYTGPFCSNEISAYFGSGSSVIYNFQENY
```

```
        ::::.::.:::::.::::.  :  :::.:::: ::::::::::::..:::. :.:::.:
gi|163 AGHCSTYGHLCRNGGRCREKRRGVTCDCAFSAYDGPFCSNEISAYFATGSSMTYHFQEHY
          830       840       850       860       870       880

            1030      1040      1050      1060      1070      1080
gi|184 LLSKNSSSHAASFHGDMKLSREMIKFSFRTTRTPSLLLFVSSFYKEYLSVIIAKNGSLQI
        ::.:::: ..:.: :. :.:::: .::::::::::::.:::::.::::::.:.::::::
gi|163 TLSENSSSLVSSLHRDVTLTREMITLSFRTTRTPSLLLYVSSFYEEYLSVILANNGSLQI
          890       900       910       920       930       940

            1090      1100      1110      1120      1130      1140
gi|184 RYKLNKYQEPDVVNFDFKNMADGQLHHIMINREEGVVFIEIDDNRRRQVHLSSGTEFSAV
       ::::...:.::. .::::::::::::.. :::::.::..:.... ..:: :::::::.::
gi|163 RYKLDRHQNPDAFTFDFKNMADGQLHQVKINREEAVVMVEVNQSTKKQVILSSGTEFNAV
          950       960       970       980       990      1000

            1150      1160      1170      1180      1190      1200
gi|184 KSLVLGRILEHSDVDQETALAGAQGFTGCLSAVQLSHVAPLKAALHPSHPDPVTVTGHVT
       :::.::..:: . .: .:  :...:::::::::.....:::::::.:: :. ::: :::.
gi|163 KSLILGKVLEAAGADPDTRRAATSGFTGCLSAVRFGRAAPLKAALRPSGPSRVTVRGHVA
         1010      1020      1030      1040      1050      1060

             1210         1220      1230      1240      1250
gi|184 E-SSCMAQPGTDATSRE---RTHSFADHSGTIDDREPLANAIKSDSAVIGGLIAVVIFIL
         . : :  .. . .::   :  . : .::  :. :::.:: . :::::::.::::::::
gi|163 PMARCAAGAASGSPARELAPRLAGGAGRSGPADEGEPLVNADRRDSAVIGGVIAVVIFIL
         1070      1080      1090      1100      1110      1120

      1260      1270      1280      1290      1300      1310
gi|184 LCITAIAVRIYQQKRLYKRSEAKRSENVDSAEAVLKSELNIQNAVNENQKEYFF
       :::::::.:::::..: :..:.: :..
gi|163 LCITAIAIRIYQQRKLRKENESKVSKKEEC
         1130      1140      1150


1311 residues in 1 query   sequences
1154 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:31:46 2002 done: Mon Dec 16 15:31:47 2002
 Scan time:  0.050 Display time:  2.066

Function used was FASTA
```

6

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAG7ai6z: 1477 aa
 >gi|14149613|ref|NP_004792.1| neurexin 1 isoform alpha precursor; neurexin I; simil
 vs  /tmp/fastaHAAH7ai6z library
searching /tmp/fastaHAAH7ai6z library

   1331 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.033
The best scores are:                                      opt
gi|6694278|gb|AAF25199.1|AF193613_1 cell recognit  (1331)  376

>>gi|6694278|gb|AAF25199.1|AF193613_1 cell recognition m  (1331 aa)
 initn: 397 init1: 163 opt: 376
Smith-Waterman score: 930;  24.015% identity in 1295 aa overlap (259-1469:165-1327)

      230       240       250       260       270       280
gi|141 LNGGVCSVVDDQAVCDCSRTGFRGKDCSQEDNNVEGLAHLMMGDQGKSKGKEEYIATFKG
                                     : : ::   : .   : :   .  . .: :
gi|669 AFPGNINSDGVVRHELQHPIIARYVRIVPLDWNGEGRIGLRIEVYGCSYWAD--VINFDG
          140       150       160       170       180         190

      290       300       310       320        330       340
gi|141 SEYFCYDLSQNPIQSSSDEITLSFKTLQRNGLMLH-TGKSADYVNLALKNGAVSLVINLG
          . : . .. ... .: :.:.::: . .:..::  :...::..: ::.. . : .:::
gi|669 HVVLPYRFRNKKMKTLKDVIALNFKTSESEGVILHGEGQQGDYITLELKKAKLVLSLNLG
            200       210       220       230       240       250

          350       360        370       380       390       400
gi|141 S---GAFEALVEPVNGKF-NDNAWHDVKVTRNLRQHSGIGHAMVTISVDGILTTTGYTQE
       :   : . . .  ..:.. .:. ::.: . :. :.        .....:  .     :.
gi|669 SNQLGPIYGHTSVMTGSLLDDHHWHSVVIERQGRS--------INLTLDRSMQHF-RTNG
            260       270       280               290        300

           410       420       430       440       450       460
gi|141 DYTMLGSDDFFYVGGSPSTADLPGSPVSNNFMGCLKEVVYKNNDVRLELSRLAKQGDPKM
       .. .:  :  .  :: : ..  :.:   .:: ::.. . :.. ..  .:.:  :. .:.
gi|669 EFDYLDLDYEITFGGIPFSGK-PSSSSRKNFKGCMESINYNGVNIT-DLAR-RKKLEPSN
           310       320        330       340       350        360

           470       480       490       500       510       520
gi|141 KIHGVVAFKCENVATLDPITFETPESFISLPKWNAKKTGSISFDFRTTEPNGLILFSH--
          : ..:.: .  :. :. :..  :.. .:   . .  :.::.::: .::::..:::
gi|669 V--GNLSFSCVEPYTV-PVFFNAT-SYLEVPGRLNQDLFSVSFQFRTWNPNGLLVFSHFA
                370        380        390       400       410

              530        540       550       560       570
gi|141 -GKPRHQKDAKHPQM-IKVDFFAIEMLDGHLYLLLDMGSGTIKIKALLKKVNDGEWYHVD
        .    . :  . .. .....   .: .      .:..::.         .:::.:..:
gi|669 DNLGNVEIDLTESKVGVHINITQTKMSQ------IDISSGS--------GLNDGQWHEVR
        420       430       440             450               460

     580       590       600       610       620       630
gi|141 FQRDGRSGTISVNTLRTPYTAPGESEILDLDDELYLGGLPENKAGLVFPTEVWTALLNYG
```

```
       :      . ....  ..  .  .    .   .. ..::. ..  .         ..:. .
gi|669 FLAKENFAILTIDGDEASAVRTNSPLQVKTGEKYFFGGFLNQMNNSSH------SVLQPS
            470       480       490       500       510

     640       650       660           670        680       690
gi|141 YVGCIRDLFIDGQSKDIRQMAEVQ--STAGVKPS-CSKETAKPCLSNPCKNNGMCRDGWN
       . ::.. . .: :  .. ..:. .  : :.:. . :.      :. : :...: : . :.
gi|669 FQGCMQLIQVDDQLVNLYEVAQRKPGSFANVSIDMCA--IIDRCVPNHCEHGGKCSQTWD
        520       530       540       550         560       570

        700       710          720                            730
gi|141 RYVCDCSGTGYLGRSCE---REATVLSY--------------DGS-------MFMKIQLP
        . : :. ::: : .:.    : .  .:              :::       ..  ..
gi|669 SFKCTCDETGYSGATCHNSIYEPSCEAYKHLGQTSNYYWIDPDGSGPLGPLKVYCNMTED
          580       590       600       610       620       630

            740           750         760         770       780
gi|141 VVMHTEAEDVSLRF----RSQRAYGI--LMATTSRD--SADTLRLELDAGRVKLTVNLDC
        :    ..:....      . . :..  :. ..: :  :: :   :     :.   ...
gi|669 KVWTIVSHDLQMQTPVVGYNPEKYSVTQLVYSASMDQISAITDSAEYCEQYVSYFCKMS-
          640       650       660       670       680       690

          790       800       810       820       830       840
gi|141 IRINCNSSKGPETLFAGYNLNDNEWHTVRVVRRGKSLKLTVDDQQAMTGQMAGDHTRLEF
        :.  . . .: : ..:   . :: :       : ..      :.   :   .      .
gi|669 -RLLNTPDGSPYTWWVG---KANEKH-YYWGGSGPGI------QKCACGIERNCTDPKYY
            700          710        720             730       740

          850       860       870       880          890       900
gi|141 HNIETGIITERRYLSSVPSNFIGHLQSLTFNGMAYIDLCKNGD---IDYCELNARFGFRN
        : ..     :.      ..:... . :  . ..  :  ..:.   ..   :  . : ::
gi|669 CNCDADYKQWRK-----DAGFLSYKDHLPVSQVVVGDTDRQGSEAKLSVGPLRCQ-GDRN
            750            760       770       780       790

             910       920       930       940       950       960
gi|141 IIADPVTFKTKSSYVALATLQAYTSMHLFFQFKTTSLDGLILYNSGDGNDFIVVELVKGY
          . ..: . :::. ..:.:. ::  . : ::: .  :..: : :  .::: .:: ..
gi|669 YW-NAASFPNPSSYLHFSTFQGETSADISFYFKTLTPWGVFLENMGK-EDFIKLELKSAT
         800       810       820       830       840        850

              970       980       990        1000      1010
gi|141 -LHYVFDLGNGANLIKGSSNKPLNDNQWHNVMISRDT--SNLHTVKIDTKITTQITAGAR
        . . ::.:::   :   :  ::::.::: :   :..  ..:.. ..  .:    : :
gi|669 EVSFSFDVGNGPVEIVVRSPTPLNDDQWHRVTAERNVKQASLQVDRLPQQIRKAPTEGHT
          860       870       880       890       900       910

     1020      1030      1040      1050      1060      1070
gi|141 NLDLKSDLYIGGVAKETYKSLPKLVHAKEGFQGCLASVDLNGRLPDLISDALFCNGQIER
        :.: :.:..::..            ...:: ::. :. .::   ::   :   .: :
gi|669 RLELYSQLFVGGAG------------GQQGFLGCIRSLRMNGVTLDLEERAKVTSGFIS-
          920                   930       940       950       960

     1080      1090      1100      1110      1120       1130
gi|141 GCEGPSTTCQEDSCSNQGVCLQQWDGFSCDCSMTSFSGPLCN-DPGTTYIFSKGGG-QIT
       :: :  :.   . : : : ::... :.::::: :...: .:: : :.  .: .:   . .
gi|669 GCSGHCTSYGTN-CENGGKCLERYHGYSCDCSNTAYDGTFCNKDVGA--FFEEGMWLRYN
             970        980       990      1000        1010

       1140                       1150      1160      1170
gi|141 YKWP--------------P---NDRPSTRADRLAIGFSTVQKEAVLVRVDSSSGLGDYLE
```

```
       .. :              :   :..:.   ... ..:::..   .:. .  ::   :.:
gi|669 FQAPATNARDSSSRVDNAPDQQNSHPDLAQEEIRFSFSTTKAPCILLYI--SSFTTDFLA
     1020      1030      1040      1050      1060        1070

    1180       1190        1200      1210      1220      1230
gi|141 LHIHQ-GKIGVKFNVG--TDDIAIEESNAIINDGKYHVVRFTRSGGNATLQVDSWPVIER
       . ..  :.. ...:.:   .   :. ..  . .:. : : .::   .  :..: .: .
gi|669 VLVKPTGSLQIRYNLGGTREPYNIDVDHRNMANGQPHSVNITRHEKTIFLKLDHYPSVSY
       1080      1090      1100      1110      1120      1130

       1240      1250                 1260       1270      1280
gi|141 YPAGRQLTIFNSQATIIIG-----GK------EQGQP-FQGQLSGLYYNGLKVLNMAAEN
       .  . . :.:::  ....:     ::      . . : : : :: . .: .  :. ::
gi|669 HLPSSSDTLFNSPKSLFLGKVIETGKIDQEIHKYNTPGFTGCLSRVQFNQIAPLK-AALR
       1140      1150      1160      1170      1180      1190

         1290      1300      1310      1320      1330      1340
gi|141 DANIAIVGNVRLVGEVPSSMTTESTATAMQSEMSTSIMETTTTLATSTARRGKPPTKEPI
       ..: .  ..:.. ::.     .::.  :                        .:  :  :.
gi|669 QTNAS--AHVHIQGEL-----VESNCGA------------------------SPLTLSPM
        1200             1210                              1220

         1350      1360      1370      1380      1390      1400
gi|141 SQTTDDILVASAECPSDDEDIDPCEPSSGGLANPTRAGGREPYPGSAEVIRESSSTTGMV
       :..::   .   .  : :   .:      : ..  : :           . ..:.  : :
gi|669 SSATDPWHLDHLDSASADFPYNP------GQGQAIRNG-----------VNRNSAIIGGV
         1230      1240            1250                 1260

         1410      1420      1430      1440      1450      1460
gi|141 VGIVAAAALCILILLYAMYKYRNRDEGSYHVDESRNYISNSAQSNGAVVKEKQPSSAKSS
       :..:  . :: :..:    .:  :  .:.::..:...  ..::.:  :..  ...:.  ...
gi|669 IAVVIFTILCTLVFLI---RYMFRHKGTYHTNEAKG--AESAESADAAIMNNDPNFTETI
      1270      1280         1290      1300        1310      1320

         1470
gi|141 NKNKKNKDKEYYV
       ...::.
gi|669 DESKKEWLI
           1330


1477 residues in 1 query   sequences
1331 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 16:01:21 2002 done: Mon Dec 16 16:01:22 2002
 Scan time:  0.033 Display time:  2.417

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAyHaq6z: 1477 aa
 >gi|14149613|ref|NP_004792.1| neurexin 1 isoform alpha precursor; neurexin I; simil
 vs  /tmp/fastaHAAzHaq6z library
searching /tmp/fastaHAAzHaq6z library

   1154 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                          opt
gi|16306509|ref|NP_387504.1| cell recognition mol   (1154)   281

>>gi|16306509|ref|NP_387504.1| cell recognition molecule   (1154 aa)
 initn: 204 init1: 152 opt: 281
Smith-Waterman score: 641;  22.746% identity in 1187 aa overlap (283-1437:183-1150)

            260       270       280       290       300       310
gi|141 KDCSQEDNNVEGLAHLMMGDQGKSKGKEEYIATFKGSEYFCYDLSQNPIQSSSDEITLSF
                                     .. : :.  . : :...:..   : :.:.:
gi|163 RFLRFLPLAWNPRGRIGMRIEVYGCAYKSEVVYFDGQSALLYRLDKKPLKPIRDVISLKF
            160       170       180       190       200       210

            320        330       340       350          360
gi|141 KTLQRNGLMLHT-GKSADYVNLALKNGAVSLVINLGSGAFEALVEPVN---GKF-NDNAW
       :..: ::..::  :. .....: : .: . . .: :.. . . . ::.   :.. .:. :
gi|163 KAMQSNGILLHREGQHGNHITLELIKGKLVFFLNSGNAKLPSTIAPVTLTLGSLLDDQHW
            220       230       240       250       260       270

       370       380       390       400       410       420
gi|141 HDVKVTRNLRQHSGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPG
       :.: .         .  ..:...::   :    .. : ..:  .  .  :: :.    ::
gi|163 HSVLIE--------LLDTQVNFTVDK-HTHHFQAKGDSSYLDLNFEISFGGIPT----PG
                    280       290        300       310

       430          440       450       460       470       480
gi|141 SPVS---NNFMGCLKEVVYKNNDVRLELSRLAKQGDPKMKIHGVVAFKCENVATLDPITF
          .   ..: :::... :.. ::    ..:::.  :.. . : :.:.: .  :. :.::
gi|163 RSRAFRRKSFHGCLENLYYNGVDV----TELAKKHKPQILMMGNVSFSCPQPQTV-PVTF
     320       330       340           350       360       370

          490       500       510       520       530       540
gi|141 ETPESFISLPKWNAKKTGSISFDFRTTEPNGLILFSHGKPRHQKDAKHPQMIKVDFFAIE
        . .:...::  ...   :..:.::: .  : .::  :. :. . .          :..
gi|163 LSSRSYLALPGNSGEDKVSVTFQFRTWNRAGHLLF--GELRRGSGS----------FVLF
          380       390       400         410                 420

          550        560       570       580       590       600
gi|141 MLDGHLYL-LLDMGSGTIKIKALLKKVNDGEWYHVDFQRDGRSGTISVNTLRTPYTAPGE
       . ::.: : :.. :..  .. :    .:::.:. :.:.      .. :.      . :
gi|163 LKDGKLKLSLFQPGQSPRNVTAG-AGLNDGQWHSVSFSAKWSHMNVVVDD--DTAVQPLV
            430       440        450       460       470

           610       620       630       640       650       660
gi|141 SEILDLDDELYLGGLPENKAGLVFPTEVWTALLNYGYVGCIRDLFIDGQSKDIRQMAEVQ
```

```
       . ..:  :  :.:              .::.. . :  .       .:. ..  ..: .
gi|163 AVLIDSGDTYYFGD------------AAWTVVQHGGPDAVTLRGAPSGHPRSAVSFAYAA
     480       490                   500       510       520

           670       680       690       700          710       720
gi|141 STAGVKPSCSKETAKPCLSNPCKNNGMCRDGWNRYVCDCSGTG---YLGRSCEREATVLS
       ... .. . .       :.. :..    : :  :   . .::    ..::. : .    .
gi|163 GAGQLRSAVN-------LAERCEQRLALRCGTARRPDSRDGTPLSWWVGRTNETH----T
       530              540       550       560       570

              730       740       750       760       770       780
gi|141 YDGSMFMKIQLPVVMHTEAEDVSLRFRSQRAYGILMATTSRDSADTLRLELDAGRVKLTV
       : :.     .::                                       :: .  . .
gi|163 YWGG-----SLP---------------------------------------DAQKCTCGL
        580                                                  590

                790       800       810       820       830
gi|141 NLDCI--RINCNSSKGPETLFAGYNLNDNEWHTVRVVRRGKSLKLTVDDQQAMT--GQMA
       . .::  .  :: . :            ::: .  .:   :  .: :  : .::  ::
gi|163 EGNCIDSQYYCNCDAG-----------RNEWTSDTIVLSQKE-HLPVT-QIVMTDTGQ--
            600                  610       620         630

        840       850       860       870       880       890
gi|141 GDHTRLEFHNIETGIITERRYLSSVPSNFIGHLQSLTFNGMAYIDLCKNGDIDYCELNAR
         :.. ..                     .: :            ::. :: ..
gi|163 -PHSEADYT--------------------LGPL------------LCR-GDQSF------
        640                                       650

        900       910       920       930       940       950
gi|141 FGFRNIIADPVTFKTKSSYVALATLQAYTSMHLFFQFKTTSLDGLILYNSGDGNDFIVVE
               . ..:.:..::. . ....  .  . : ::::  .:... : :  .::: .:
gi|163 -------WNSASFNTETSYLHFPAFHGELTADVCFFFKTTVSSGVFMENLGI-TDFIRIE
              660       670       680       690       700

         960       970       980       990        1000      1010
gi|141 L-VKGYLHYVFDLGNGANLIKGSSNKPLNDNQWHNVMISRDT--SNLHTVKIDTKITTQI
       : .   . . ::.:::   .  .:  :.::::::.:   :..  ..:.. ..  :.
gi|163 LRAPTEVTFSFDVGNGPCEVTVQSPTPFNDNQWHHVRAERNVKGASLQVDQLPQKMQPAP
     710       720       730       740       750       760

          1020      1030      1040      1050      1060      1070
gi|141 TAGARNLDLKSDLYIGGVAKETYKSLPKLVHAKEGFQGCLASVDLNGRLPDLISDALFCN
       . :   :.:.:.:.:::.: .           ..:: ::. :..:::   ::   :
gi|163 ADGHVRLQLNSQLFIGGTATR-----------QRGFLGCIRSLQLNGVALDLEERATVTP
     770       780       790                  800       810

          1080      1090      1100      1110      1120      1130
gi|141 GQIERGCEGPSTTCQEDSCSNQGVCLQQWDGFSCDCSMTSFSGPLCNDPGTTYIFSKGGG
       : .: :: :  .:  .  : : : : ..  : .:::......::.:..  ..:. .  :.
gi|163 G-VEPGCAGHCSTYGH-LCRNGGRCREKRRGVTCDCAFSAYDGPFCSNEISAYFAT--GS
       820       830        840       850       860       870

          1140      1150      1160      1170      1180       1190
gi|141 QITYKWPPNDRPSTRADRLAIGFSTVQKEAVLVRVDSSSGLGDYLELHIHQGKI-GVKFN
       ..::..  .   :  .. :.   :.......:.:        ... : ..  .  .. .
gi|163 SMTYHFQEHYTLSENSSSLV---SSLHRDVTLTR--------EMITLSFRTTRTPSLLLY
          880       890          900               910       920

           1200        1210       1220      1230      1240
gi|141 VGTDDIAIEESNAII--NDGKYHV-VRFTRSGGNATLQVDSWPVIERYPAGR--QLTIFN
```

```
       :..     ::  ..:  :.:. ..  .. :  .  ..  :    .. .  :.  :. :
gi|163 VSS---FYEEYLSVILANNGSLQIRYKLDRHQNPDAFTFD----FKNMADGQLHQVKINR
              930       940       950       960           970

      1250      1260         1270      1280      1290       1300
gi|141 SQATIIIGGKEQGQPFQGQL---SGLYYNGLKVLNMAAENDANIAIVGNV-RLVGEVPSS
        .:....   : .:  . :.   ::  .:..: :           :.:.: . .:  :.
gi|163 EEAVVMV---EVNQSTKKQVILSSGTEFNAVKSL-----------ILGKVLEAAGADPD-
       980           990      1000                 1010      1020

          1310      1320      1330      1340      1350        1360
gi|141 MTTESTATAMQSEMSTSIMETTTTLATSTARRGKPPTKEPISQTTDDI--LVASAECPSD
        : ....... . .:.  .  .. :  ..: : . :..  .   .  .   .:.:   :
gi|163 -TRRAATSGFTGCLSAVRFGRAAPL--KAALRPSGPSRVTVRGHVAPMARCAAGAASGSP
             1030      1040        1050      1060      1070

            1370      1380      1390      1400      1410      1420
gi|141 DEDIDPCEPSSGGLANPTRAGGREPYPGSAEVIRESSSTTGMVVGIVAAAALCILILLYA
        ... :   ...: ..:.  :  ::  ..    :..:.. : :...:    :::  .
gi|163 ARELAPRLAGGAGRSGPADEG--EPLVNAD---RRDSAVIGGVIAVVIFILLCITAIAIR
     1080      1090        1100         1110      1120      1130

             1430      1440      1450      1460      1470
gi|141 MYKYRN-RDEGSYHVDESRNYISNSAQSNGAVVKEKQPSSAKSSNKNKKNKDKEYYV
       .:. :. : :.  .:...
gi|163 IYQQRKLRKENESKVSKKEEC
          1140      1150


1477 residues in 1 query   sequences
1154 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 16:02:13 2002 done: Mon Dec 16 16:02:14 2002
 Scan time:  0.050 Display time:  1.967

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaKAACHaq6z: 1477 aa
 >gi|14149613|ref|NP_004792.1| neurexin 1 isoform alpha precursor; neurexin I; simil
 vs  /tmp/fastaLAADHaq6z library
searching /tmp/fastaLAADHaq6z library

   1311 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.017
The best scores are:                                      opt
gi|18496979|ref|NP_207837.1| cell recognition pro  (1311)  403

>>gi|18496979|ref|NP_207837.1| cell recognition protein   (1311 aa)
 initn: 265 init1: 150 opt: 403
Smith-Waterman score: 866;  25.392% identity in 1020 aa overlap (283-1229:186-1123)

            260       270       280       290       300       310
gi|141 KDCSQEDNNVEGLAHLMMGDQGKSKGKEEYIATFKGSEYFCYDLSQNPIQSSSDEITLSF
                                     .. . :.  . : ..:. ..  .: :.:.:
gi|184 RFLRFIPLEWNPKGRIGMRIEVFGCAYRSEVVDLDGKSSLLYRFDQKSLSPIKDIISLKF
         160       170       180       190       200       210

            320        330       340       350          360
gi|141 KTLQRNGLMLHT-GKSADYVNLALKNGAVSLVINLGSGAFEALVEPVN---GKF-NDNAW
       ::.: .:..::  : ..:...: :. . . :.:: : . . .    ::   :.. .:. :
gi|184 KTMQSDGILLHREGPNGDHITLQLRRARLFLLINSGEAKLPSTSTLVNLTLGSLLDDQHW
         220       230       240       250       260       270

       370       380       390       400       410       420
gi|141 HDVKVTRNLRQHSGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPG
       :.: . :  .:        :...::       ... ...... :  .  :: :.    ::
gi|184 HSVLIQRLGKQ--------VNFTVDE-HRHHFHARGEFNLMNLDYEISFGGIPA----PG
         280               290        300       310       320

       430          440       450       460       470       480
gi|141 SPVS---NNFMGCLKEVVYKNNDVRLELSRLAKQGDPKMKIHGVVAFKCENVATLDPITF
       . ::    :: :::... :.. :.      ::::  :..   : :.:.: .  .. :.::
gi|184 KSVSFPHRNFHGCLENLYYNGVDI----IDLAKQQKPQIIAMGNVSFSCSQPQSM-PVTF
            330       340           350       360       370

          490       500       510       520       530       540
gi|141 ETPESFISLPKWNAKKTGSISFDFRTTEPNGLILFSHGKPRHQKDAKHPQMIKVDFFAIE
        . .:...:: .....  : .:.::: .  ::.:::.            :.:.  .. .
gi|184 LSSRSYLALPDFSGEEEVSATFQFRTWNKAGLLLFSEL-----------QLISGGIL-LF
       380       390       400       410                  420

          550        560       570       580       590       600
gi|141 MLDGHLYL-LLDMGSGTIKIKALLKKVNDGEWYHVDFQRDGRSGTISVNTLRTPYTAP--
       . ::.:   : . :.    : : .. .:::.:. :...      ...:.  .   .::
gi|184 LSDGKLKSNLYQPGKLPSDITAGVE-LNDGQWHSVSLSAKKNHLSVAVDG-QMASAAPLL
         430       440       450        460       470        480

             610       620       630       640       650
gi|141 GESEILDLDDELYLGGLPENKAGLVFPTEVWTALLNYGYVGCIRDLFIDGQSKDI--RQM
```

```
       :  .: .     :.:: :...    : ..  . :   :. ::.: . :.:.  :.   :.
gi|184 GPEQIYS-GGTYYFGGCPDKS----FGSKCKSPL--GGFQGCMRLISISGKVVDLISVQQ
           490        500           510         520       530

     660        670       680       690       700       710
gi|141 AEVQSTAGVK-PSCSKETAKPCLSNPCKNNGMCRDGWNRYVCDCSGTGYLGRSC-----E
       . . . . ..  ::.   .  :: : :...: : ..:. . :.:..::: : .:     :
gi|184 GSLGNFSDLQIDSCG--ISDRCLPNYCEHGGECSQSWSTFHCNCTNTGYRGATCHNSIYE
        540       550         560       570       580       590

           720         730       740       750       760       770
gi|141 REATVLSYDG--SMFMKIQLPVVMHTEAEDVSLRFRSQRAYGILMATTSRDSADTLRLEL
       .   .. . :  : :. :.       :   .   . .. :. :..    ....:. :..
gi|184 QSCEAYKHRGNTSGFYYIDSDGSGPLEPFLLYCNM-TETAWTIIQ----HNGSDLTRVRN
          600       610       620        630           640

                   780       790       800                 810
gi|141 D------AGRVKLTVNLDCIRINCNSSKGPETLFAGY-------NLNDN---EWHTVRVV
              ::  . ..... .. . : ..  :  :. :       : .:.    : . :.
gi|184 TNPENPYAGFFEYVASMEQLQATINRAEHCEQEFTYYCKKSRLVNKQDGTPLSWWVGRTN
     650       660       670       680       690       700

         820       830                  840       850         860
gi|141 RRGKSLKLTVDDQQAMTGQMAG-----------DHTRLEFHNIETGIITERRYL--SSVP
       .       .  : :  :  . :           :  : :. : .::... ...:  ...
gi|184 ETQTYWGGSSPDLQKCTCGLEGNCIDSQYYCNCDADRNEWTN-DTGLLAYKEHLPVTKIV
     710       720       730       740       750        760

            870       880       890       900       910       920
gi|141 SNFIGHLQSLTFNGMAYIDLCKNGDIDYCELNARFGFRNIIADPVTFKTKSSYVALATLQ
        .  :.:.: .   .. . ::. :: ..              . ..: :..::. . :..
gi|184 ITDTGRLHSEAAYKLGPL-LCR-GDRSF-------------WNSASFDTEASYLHFPTFH
      770       780         790                    800       810

            930       940       950       960        970       980
gi|141 AYTSMHLFFQFKTTSLDGLILYNSGDGNDFIVVELVKG-YLHYVFDLGNGANLIKGSSNK
       .  :  . : ::::. .:..: : : . ::: .:: .   . . ::.:::   :. .:
gi|184 GELSADVSFFFKTTASSGVFLENLGIA-DFIRIELRSPTVVTFSFDVGNGPFEISVQSPT
           820       830       840        850       860       870

             990        1000      1010      1020      1030
gi|141 PLNDNQWHNVMISRD--TSNLHTVKIDTKITTQITAGARNLDLKSDLYIGGVAKETYKSL
        .::::::.: . :.   ..:.. ..  :     . :   :.:.:.:..::.: .
gi|184 HFNDNQWHHVRVERNMKEASLQVDQLTPKTQPAPADGHVLLQLNSQLFVGGTATR-----
            880       890       900       910       920

    1040      1050      1060      1070      1080      1090
gi|141 PKLVHAKEGFQGCLASVDLNGRLPDLISDALFCNGQIERGCEGPSTTCQEDSCSNQGVCL
             ..:: ::. :..:::   ::   :   . ... ::.:  ..  .  : : : :
gi|184 ------QRGFLGCIRSLQLNGMTLDLEERAQV-TPEVQPGCRGHCSSYGK-LCRNGGKCR
             930       940       950        960       970

    1100      1110      1120      1130      1140
gi|141 QQWDGFSCDCSMTSFSGPLCNDPGTTYIFSKGGGQITYKWPPN----DRPSTRA------
       ..  :: :::......::.:..  ..:. :  :... :..  :       :..:
gi|184 ERPIGFFCDCTFSAYTGPFCSNEISAYFGS--GSSVIYNFQENYLLSKNSSSHAASFHGD
     980       990      1000        1010      1020      1030

       1150        1160      1170      1180       1190
gi|141 ---DRLAIGFS--TVQKEAVLVRVDSSSGLGDYLELHI-HQGKIGVKFNVGT----DDIA
```

```
          .:  : ::  :..  ..:. :  ::   .:: . : ..:.. ......     : .
gi|184 MKLSREMIKFSFRTTRTPSLLLFV--SSFYKEYLSVIIAKNGSLQIRYKLNKYQEPDVVN
      1040      1050      1060        1070      1080      1090

    1200      1210      1220      1230      1240      1250
gi|141 IEESNAIINDGKYHVVRFTRSGGNATLQVDSWPVIERYPAGRQLTIFNSQATIIIGGKEQ
       .. .:  . ::. : . ..:  : . ...:.
gi|184 FDFKN--MADGQLHHIMINREEGVVFIEIDDNRRRQVHLSSGTEFSAVKSLVLGRILEHS
        1100        1110      1120      1130      1140      1150


1477 residues in 1 query   sequences
1311 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 16:03:13 2002 done: Mon Dec 16 16:03:14 2002
 Scan time:  0.017 Display time:  1.850

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAsxa4Wh: 1642 aa
 >gi|21166380|ref|NP_620060.1| neurexin 2, isoform alpha-2 precursor; neurexin II [H
 vs  /tmp/fastaHAAtxa4Wh library
searching /tmp/fastaHAAtxa4Wh library

   1331 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 41, opt: 29, gap-pen: -12/ -2, width:  16
 Scan time:  0.066
The best scores are:                                           opt
gi|6694278|gb|AAF25199.1|AF193613_1 cell recognit  (1331)  326

>>gi|6694278|gb|AAF25199.1|AF193613_1 cell recognition m  (1331 aa)
 initn: 434 init1: 185 opt: 326
Smith-Waterman score: 807;  23.942% identity in 1111 aa overlap (265-1285:187-1228)

          240       250       260       270       280       290
gi|211 GFGGKFCSEEEHPMEGPAHLTLNSEGKEEFVATFKGNEFFCYDLSHNPIQSSTDEITLAF
                                     : .: :.  . : . .. ...  : :.: :
gi|669 RYVRIVPLDWNGEGRIGLRIEVYGCSYWADVINFDGHVVLPYRFRNKKMKTLKDVIALNF
        160       170       180       190       200       210

          300       310       320       330       340
gi|211 RTLQRNGLMLH-TGKSADYVNLSLKSGAVWLVINLGS---GAFEALVEPVNGKF-NDNAW
       .: . .:..::  :...::..: ::.. . : .::::   : . . .  ..:.. .:. :
gi|669 KTSESEGVILHGEGQQGDYITLELKKAKLVLSLNLGSNQLGPIYGHTSVMTGSLLDDHHW
        220       230       240       250       260       270

     350       360       370       380       390       400
gi|211 HDVRVTRNLRQHAGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYIGGSPNTADLPG
       :.: . :. :.        .....:  .     :. .. .:  :  . .:: : ..  :.
gi|669 HSVVIERQGRS--------INLTLDRSMQHF-RTNGEFDYLDLDYEITFGGIPFSGK-PS
        280               290        300       310       320

     410       420       430       440       450       460
gi|211 SPVSNNFMGCLKDVVYKNNDFKLELSRLAKEGDPKMKLQGDLSFRCEDVAALDPVTFESP
       :   .:: ::.... :.. ..  .:.:  :. .:.    :.::: : .  .. :: :..
gi|669 SSSRKNFKGCMESINYNGVNIT-DLAR-RKKLEPSNV--GNLSFSCVEPYTV-PVFFNAT
        330       340        350        360         370        380

     470       480       490       500       510       520
gi|211 EAFVALPRWSAKRTGSISLDFRTTEPNGLLLFSQGRRAGGGAGSHSSAQRADYFAMELLD
        ... .:    .   :.:..::: .:::::.::.     :..    . ...    ... .
gi|669 -SYLEVPGRLNQDLFSVSFQFRTWNPNGLLVFSHFADNLGNVEIDLTESKVGVH-INITQ
              390       400       410       420       430

     530       540       550       560       570       580
gi|211 GHLYLLLDMGSGGIKLRASSRKVNDGEWCHVDFQRDGRKGSISVNSRSTPFLATGDSEIL
        ..  . :..::.         .:::.: .: :      . .....  .  . :..   .
gi|669 TKMSQI-DISSGS--------GLNDGQWHEVRFLAKENFAILTIDGDEASAVRTNSPLQV
     440        450               460       470       480       490

     590       600       610       620       630       640
gi|211 DLESELYLGGLPEGGRVDLPLPPEVWTAALRAGYVGCVRDLFIDGRSRDLRGLAEAQ-GA
```

```
           . ..::.       :    .   ..:. .. ::.. . .: .  .:   .:. . :.
gi|669 KTGEKYFFGGF-------LNQMNNSSHSVLQPSFQGCMQLIQVDDQLVNLYEVAQRKPGS
              500              510       520       530       540

      650       660       670       680       690          700
gi|211 VGVAPFCSRETLKQCASAPCRNGGVCREGWNRFICDCIGTGFLGRVCER---EATVLSY-
        . . .     . .:.   :..:: : . :. : : :  ::. : .:.    : .  .:
gi|669 FANVSIDMCAIIDRCVPNHCEHGGKCSQTWDSFKCTCDETGYSGATCHNSIYEPSCEAYK
           550       560       570       580       590       600

                              710       720                  730
gi|211 -------------DGS-------MYMKIMLPNAMHTEAEDVSLRF-----------MSQR
                    :::       .: ..   ..    ..:....            ..:
gi|669 HLGQTSNYYWIDPDGSGPLGPLKVYCNMTEDKVWTIVSHDLQMQTPVVGYNPEKYSVTQL
           610       620       630       640       650       660

           740         750        760        770       780
gi|211 AYGLMMATTS--RESADTLRLELDG-GQMKLTVNLGKG-PETLFAGHKLNDNEWHTVRVV
       .:.  :   :   .::.  .  ..   .:.  .:   : : : ..: : :.....
gi|669 VYSASMDQISAITDSAEYCEQYVSYFCKMSRLLNTPDGSPYTWWVG-KANEKHYYWGG--
           670       680       690       700        710       720

     790       800       810       820       830       840
gi|211 RRGKSLQLSVDNVTVEGQMAGAHMRLEFHNIETGIMTERRFISVVPSNFIGHL--SGLVF
         : ..:  .    .: . .  ..   . : ..     :.  . .  ..  ::  : .:
gi|669 -SGPGIQKCA--CGIERNCTDPKY---YCNCDADYKQWRKDAGFL--SYKDHLPVSQVVV
                 730       740          750         760       770

       850       860       870       880       890       900
gi|211 NGQPYMDQCKDGDITYCELNARFGLRAIVADPVTFKSRSSYLALATLQAYASMHLFFQFK
       .      : ... ..   :  . : :    . ..: . :::: ..:.:. .:  . : ::
gi|669 GDTDR--QGSEAKLSVGPLRCQ-GDRNYW-NAASFPNPSSYLHFSTFQGETSADISFYFK
              780       790         800       810       820

       910       920       930        940       950       960
gi|211 TTAPDGLLLFNSGNGNDFIVIELVKGY-IHYVFDLGNGPSLMKGNSDKPVNDNQWHNVVV
       : .: :..: : :. .::: .:: ..  . . ::.::::  .   :  :.::.::: :..
gi|669 TLTPWGVFLENMGK-EDFIKLELKSATEVSFSFDVGNGPVEIVVRSPTPLNDDQWHRVTA
      830       840        850       860       870       880

          970       980       990      1000      1010      1020
gi|211 SRD--PGNVHTLKIDSRTVTQHSNGARNLDLKGELYIGGLSKNMFSNLPKLVASRDGFQG
        :.   ..... .. ..     ..:   :.: ..:..::            .....:: :
gi|669 ERNVKQASLQVDRLPQQIRKAPTEGHTRLELYSQLFVGG------------AGGQQGFLG
       890       900       910       920                   930

         1030      1040      1050      1060      1070      1080
gi|211 CLASVDLNGRLPDLIADALHRIGQVERGCDGPSTTCTEESCANQGVCLQQWDGFTCDCTM
       :. :. .::   ::   :    : .  ::.:  :.    .: : : ::... :..:::.
gi|669 CIRSLRMNGVTLDLEERAKVTSGFIS-GCSGHCTSYGT-NCENGGKCLERYHGYSCDCSN
         940       950       960        970        980       990

         1090       1100      1110                         1120
gi|211 TSYGGPVCN-DPGTTYIFGKGGALITYTW--P--------------P---NDRPSTRMDR
       :.: :  :: : :.   : ..:  . :..  :              :   :..:.  ...
gi|669 TAYDGTFCNKDVGA---FFEEGMWLRYNFQAPATNARDSSSRVDNAPDQQNSHPDLAQEE
          1000         1010      1020      1030      1040      1050

         1130      1140      1150       1160        1170      1180
gi|211 LAVGFSTHQRSAVLVRVDSASGLGDYLQLHIDQ-GTVGVIFNVG--TDDITIDEPNAIVS
```

```
       .  .::: .   .:. ..: .   :.: . .   :.. . .:.:   .  .::  .  ..
gi|669 IRFSFSTTKAPCILLYISSFTT--DFLAVLVKPTGSLQIRYNLGGTREPYNIDVDHRNMA
             1060      1070        1080      1090      1100

            1190      1200       1210      1220           1230
gi|211 DGKYHVVRFTRSGGNATLQVDSWP-VNERYPAGRQLTIFNSQAAIKIG-----GR-DQ--
       .:. : : .::   .  :..: .: :. . :.. . :.:::  .. .:     :. ::  
gi|669 NGQPHSVNITRHEKTIFLKLDHYPSVSYHLPSSSD-TLFNSPKSLFLGKVIETGKIDQEI
     1110      1120      1130      1140       1150      1160

               1240      1250         1260      1270       1280
gi|211 ---GRP-FQGQVSGLYYNGLKVLALA---AESDPNVRTEGHLRLVGEGPSVL-LSAETTA
          . : : : .: . .: .  :  :   .... .:. .:.:   . : : : ::  ..:
gi|669 HKYNTPGFTGCLSRVQFNQIAPLKAALRQTNASAHVHIQGELVESNCGASPLTLSPMSSA
      1170      1180      1190      1200      1210      1220

         1290      1300      1310      1320      1330      1340
gi|211 TTLLADMATTIMETTTTMATTTTRRGRSPTLRDSTTQNTDDLLVASAECPSDDEDLEECE
       :
gi|669 TDPWHLDHLDSASADFPYNPGQGQAIRNGVNRNSAIIGGVIAVVIFTILCTLVFLIRYMF
      1230      1240      1250      1260      1270      1280


1642 residues in 1 query   sequences
1331 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:51:21 2002 done: Mon Dec 16 15:51:22 2002
 Scan time:  0.066 Display time:  2.117

Function used was FASTA
```

K

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAuHaq6z: 1642 aa
 >gi|21166380|ref|NP_620060.1| neurexin 2, isoform alpha-2 precursor; neurexin II [H
 vs  /tmp/fastaDAAvHaq6z library
searching /tmp/fastaDAAvHaq6z library

   1154 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 41, opt: 29, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                      opt
gi|16306509|ref|NP_387504.1| cell recognition mol  (1154)  262

>>gi|16306509|ref|NP_387504.1| cell recognition molecule  (1154 aa)
 initn: 372 init1: 151 opt: 262
Smith-Waterman score: 660;  25.153% identity in 982 aa overlap (265-1202:183-986)

          240       250       260       270       280       290
gi|211 GFGGKFCSEEEHPMEGPAHLTLNSEGKEEFVATFKGNEFFCYDLSHNPIQSSTDEITLAF
                                     :. : :.  . : :...:..   : :.: :
gi|163 RFLRFLPLAWNPRGRIGMRIEVYGCAYKSEVVYFDGQSALLYRLDKKPLKPIRDVISLKF
            160       170       180       190       200       210

          300        310       320       330       340
gi|211 RTLQRNGLMLHT-GKSADYVNLSLKSGAVWLVINLGSGAFEALVEPVN---GKF-NDNAW
       ...: ::..::  :. .....: : .: . . .: :.. . . . ::.   :.. .:. :
gi|163 KAMQSNGILLHREGQHGNHITLELIKGKLVFFLNSGNAKLPSTIAPVTLTLGSLLDDQHW
            220       230       240       250       260       270

     350       360       370       380       390       400
gi|211 HDVRVTRNLRQHAGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYIGGSPNTADLPG
       :.: .         .  ..:...::   :    .. : ..:  .  . .:: :.    ::
gi|163 HSVLIE--------LLDTQVNFTVDK-HTHHFQAKGDSSYLDLNFEISFGGIPT----PG
                    280       290        300       310

     410          420       430       440       450       460
gi|211 SPVS---NNFMGCLKDVVYKNNDFKLELSRLAKEGDPKMKLQGDLSFRCEDVAALDPVTF
          .   ..: :::... :.. :    ...:::.  :.. ..:..:: : .  .. ::::
gi|163 RSRAFRRKSFHGCLENLYYNGVD----VTELAKKHKPQILMMGNVSFSCPQPQTV-PVTF
     320       330       340           350       360       370

        470       480       490       500       510       520
gi|211 ESPEAFVALPRWSAKRTGSISLDFRTTEPNGLLLFSQGRRAGGGAGSHSSAQRADYFAME
        : ....:::  :..   :....::: .  : :::.. ::   :.::         :..
gi|163 LSSRSYLALPGNSGEDKVSVTFQFRTWNRAGHLLFGELRR---GSGS---------FVLF
          380       390       400       410                   420

        530        540       550       560       570         580
gi|211 LLDGHLYL-LLDMGSGGIKLRASSRKVNDGEWCHVDFQRDGRKGSISVNSRST--PFLAT
       : ::.: : :.. :..  .. :..  .:::.:  :.:.    . .. :.. ..  :..:.
gi|163 LKDGKLKLSLFQPGQSPRNVTAGA-GLNDGQWHSVSFSAKWSHMNVVVDDDTAVQPLVAV
            430       440        450       460       470       480

           590       600       610       620       630       640
gi|211 GDSEILDLESELYLGGLPEGGRVDLPLPPEVWTAALRAGYVGCVRDLFIDGRSRDLRGLA
```

```
           ..:  .  :.:               .::.. ..:  . .     .:. :.  ..:
gi|163 ----LIDSGDTYYFGD-------------AAWTVVQHGGPDAVTLRGAPSGHPRSAVSFA
                 490                    500       510       520

                 650       660         670       680       690
gi|211 EAQGA------VGVAPFCSRETLKQCASA--PCRNGGVCREGWNRFICDCIGTGFLGRVC
        : ::      :..:  : ..   .:..:  :    :.    :            .::.
gi|163 YAAGAGQLRSAVNLAERCEQRLALRCGTARRPDSRDGTPLSWW------------VGRTN
          530       540       550       560                   570

         700       710       720       730       740       750
gi|211 EREATVLSYDGSMYMKIMLPNAMHTEAEDVSLRFMSQRAYGLMMATTSRESADTLRLELD
       :   :   . ::      ::.:..                ::       .  :.
gi|163 E---THTYWGGS------LPDAQKCTC-------------GL-----EGNCIDS------
               580             590

         760       770       780       790       800       810
gi|211 GGQMKLTVNLGKGPETLFAGHKLNDNEWHTVRVVRRGKSLQLSVDNVTVEGQMAGAHMRL
         :.  . . :.             ::: .  .:   :  .: : ..
gi|163 --QYYCNCDAGR-------------NEWTSDTIVLSQKE-HLPVTQI-------------
       600                    610       620        630

         820       830       840       850       860       870
gi|211 EFHNIETGIMTERRFISVVPSNFIGHLSGLVFNGQPYMDQCKDGDITYCELNARFGLRAI
               .::.                    .:::.    ...: :   :  : : ...
gi|163 --------VMTD--------------------TGQPH----SEADYTLGPLLCR-GDQSF
                                              640       650

         880       890       900       910       920       930
gi|211 VADPVTFKSRSSYLALATLQAYASMHLFFQFKTTAPDGLLLFNSGNGNDFIVIEL-VKGY
         . ..:....::: . ....  .  . : ::::. .:... : :  .::: ::: .
gi|163 W-NSASFNTETSYLHFPAFHGELTADVCFFFKTTVSSGVFMENLGI-TDFIRIELRAPTE
        660       670       680       690       700        710

          940       950       960       970          980       990
gi|211 IHYVFDLGNGPSLMKGNSDKPVNDNQWHNVVVSRDPGNVHTLKID---SRTVTQHSNGAR
       . . ::.::::  .  .:  : ::::::.: . :.  .. .:..:   ..     ..:
gi|163 VTFSFDVGNGPCEVTVQSPTPFNDNQWHHVRAERNVKGA-SLQVDQLPQKMQPAPADGHV
         720       730       740       750        760       770

            1000      1010      1020       1030      1040      1050
gi|211 NLDLKGELYIGGLSKNMFSNLPKLVASRD-GFQGCLASVDLNGRLPDLIADALHRIGQVE
        :.:...:.:::            .:.:. :: ::. :..:::   ::   :    : ::
gi|163 RLQLNSQLFIGG------------TATRQRGFLGCIRSLQLNGVALDLEERATVTPG-VE
          780                   790       800       810        820

             1060      1070      1080      1090      1100
gi|211 RGCDGPSTTCTEESCANQGVCLQQWDGFTCDCTMTSYGGPVCNDPGTTYIFGKGGALIT-
        :: :  .:  .  : : : :  ..  : ::::....: :: :..  ..: :. :...
gi|163 PGCAGHCSTYGHL-CRNGGRCREKRRGVTCDCAFSAYDGPFCSNEISAY-FATGSSMTYH
             830        840       850       860        870

        1110                1120      1130      1140      1150
gi|211 ----YTWPPN----------DRPSTRMDRLAVGFSTHQRSAVLVRVDSASGLGDYLQLHI
           ::   :          :   :: . ....: : .  ..:. :.:     .::.. .
gi|163 FQEHYTLSENSSSLVSSLHRDVTLTR-EMITLSFRTTRTPSLLLYVSSF--YEEYLSVIL
     880     890       900        910       920         930

         1160          1170      1180      1190      1200      1210
gi|211 -DQGTVGVIFNV----GTDDITIDEPNAIVSDGKYHVVRFTRSGGNATLQVDSWPVNERY
```

```
        ..:.. . ...    . : .:.:  :  ..::. : :...:  . . ..:..
gi|163 ANNGSLQIRYKLDRHQNPDAFTFDFKN--MADGQLHQVKINREEAVVMVEVNQSTKKQVI
        940       950       960         970       980       990

             1220      1230      1240      1250      1260      1270
gi|211 PAGRQLTIFNSQAAIKIGGRDQGRPFQGQVSGLYYNGLKVLALAAESDPNVRTEGHLRLV

gi|163 LSSGTEFNAVKSLILGKVLEAAGADPDTRRAATSGFTGCLSAVRFGRAAPLKAALRPSGP
         1000      1010      1020      1030      1040      1050


1642 residues in 1 query   sequences
1154 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:53:54 2002 done: Mon Dec 16 15:53:55 2002
 Scan time:  0.050 Display time:  1.700

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAe9aWNs: 1642 aa
 >gi|21166380|ref|NP_620060.1| neurexin 2, isoform alpha-2 precursor; neurexin II [H
 vs  /tmp/fastaHAAf9aWNs library
searching /tmp/fastaHAAf9aWNs library

   1311 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 41, opt: 29, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                              opt
gi|18496979|ref|NP_207837.1| cell recognition pro  (1311)  523

>>gi|18496979|ref|NP_207837.1| cell recognition protein   (1311 aa)
 initn: 483 init1: 170 opt: 523
Smith-Waterman score: 877;  25.980% identity in 1020 aa overlap (265-1202:186-1123)

          240       250       260       270       280       290
gi|211 GFGGKFCSEEEHPMEGPAHLTLNSEGKEEFVATFKGNEFFCYDLSHNPIQSSTDEITLAF
                                      :. . :.  . : .... ..   : :.: :
gi|184 RFLRFIPLEWNPKGRIGMRIEVFGCAYRSEVVDLDGKSSLLYRFDQKSLSPIKDIISLKF
         160       170       180       190       200       210

          300       310       320       330       340
gi|211 RTLQRNGLMLHT-GKSADYVNLSLKSGAVWLVINLGSGAFEALVEPVN---GKF-NDNAW
       .:.: .:..::  : ..:...:.:. . ..:.:: : . .    ::   :.. .:. :
gi|184 KTMQSDGILLHREGPNGDHITLQLRRARLFLLINSGEAKLPSTSTLVNLTLGSLLDDQHW
         220       230       240       250       260       270

     350       360       370       380       390       400
gi|211 HDVRVTRNLRQHAGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYIGGSPNTADLPG
       :.: . :  .:        :...::       ... ...... :  . .:: :     ::
gi|184 HSVLIQRLGKQ--------VNFTVDE-HRHHFHARGEFNLMNLDYEISFGGIPA----PG
         280               290        300       310       320

     410          420       430       440       450       460
gi|211 SPVS---NNFMGCLKDVVYKNNDFKLELSRLAKEGDPKMKLQGDLSFRCEDVAALDPVTF
       . ::    :: :::... :.. :    .  :::.  :..  .:..:: : .  .. ::::
gi|184 KSVSFPHRNFHGCLENLYYNGVD----IIDLAKQQKPQIIAMGNVSFSCSQPQSM-PVTF
            330       340           350       360       370

        470       480       490       500       510       520
gi|211 ESPEAFVALPRWSAKRTGSISLDFRTTEPNGLLLFSQGRRAGGGAGSHSSAQRADYFAME
        :  ...::: .:...  : ...::: .  ::::::. .  .::            . .
gi|184 LSSRSYLALPDFSGEEEVSATFQFRTWNKAGLLLFSELQLISGG------------ILLF
       380       390       400       410       420

        530        540       550       560       570          580
gi|211 LLDGHLYL-LLDMGSGGIKLRASSRKVNDGEWCHVDFQRDGRKGSISVNSR---STPFLA
       : ::.:   : . :.    . :.  ..:::.:  :...    . :..:...   ..:.:
gi|184 LSDGKLKSNLYQPGKLPSDITAGV-ELNDGQWHSVSLSAKKNHLSVAVDGQMASAAPLL-
         430       440        450       460       470       480

            590       600          610       620       630
gi|211 TGDSEILDLESELYLGGLPE---GGRVDLPLPPEVWTAALRAGYVGCVRDLFIDGRSRDL
```

```
        :  .: .  .  :.:: :.   :..   ::          .:. ::.: . :.:.  ::
gi|184 -GPEQIYS-GGTYYFGGCPDKSFGSKCKSPL----------GGFQGCMRLISISGKVVDL
            490        500       510                 520       530

     640       650       660          670       680       690
gi|211 RGLAEAQGAVGVAPFCSRETL---KQCASAPCRNGGVCREGWNRFICDCIGTGFLGRVCE
         ..  ::..:     . ..     .:    :..:: : ..:. : :.: .::. : .:.
gi|184 --ISVQQGSLGNFSDLQIDSCGISDRCLPNYCEHGGECSQSWSTFHCNCTNTGYRGATCH
               540       550       560       570       580

           700             710            720       730        740
gi|211 R---EATVLSY------DGSMYMKI-----MLPNAMHTEAEDVSLRFMSQRAYGLM-MAT
           : .  .:      .: .:.       . :  .. .  ...  .... .  :  . .
gi|184 NSIYEQSCEAYKHRGNTSGFYYIDSDGSGPLEPFLLYCNMTETAWTIIQHNGSDLTRVRN
     590       600       610       620       630       640

               750       760       770          780
gi|211 TSRES--ADTLRLELDGGQMKLTVNLGKGPETLFAGH----KLNDNE------WHTVRVV
       :. :.  :  ..   .  :.. :.: ..  :  :. .    .: ...      : . :.
gi|184 TNPENPYAGFFEYVASMEQLQATINRAEHCEQEFTYYCKKSRLVNKQDGTPLSWWVGRTN
     650       660       670       680       690       700

     790            800       810              820       830
gi|211 RR-----GKSLQLSVDNVTVEGQMAGAHM-------RLEFHNIETGIMTERRFISVVPSN
       .      :.: .:.  .  .::.   ...       : :. : .::... .. . :  ..
gi|184 ETQTYWGGSSPDLQKCTCGLEGNCIDSQYYCNCDADRNEWTN-DTGLLAYKEHLPV--TK
     710       720       730       740       750        760

       840       850       860       870       880       890
gi|211 FIGHLSGLVFNGQPYMDQCKDGDITYCELNARFGLRAIVADPVTFKSRSSYLALATLQAY
       ..   .: . .   :    : : .  :. .  :       . ..: ...::: . :...
gi|184 IVITDTGRLHSEAAY----KLGPLL-CRGDRSFW------NSASFDTEASYLHFPTFHGE
        770       780            790             800       810

       900       910       920       930        940       950
gi|211 ASMHLFFQFKTTAPDGLLLFNSGNGNDFIVIELVKG-YIHYVFDLGNGPSLMKGNSDKPV
        :  . : ::::: .:..: : : . ::: ::: .   . . ::.::::  .. .:
gi|184 LSADVSFFFKTTASSGVFLENLGIA-DFIRIELRSPTVVTFSFDVGNGPFEISVQSPTHF
         820       830       840        850       860       870

        960       970       980          990      1000      1010
gi|211 NDNQWHNVVVSRDPGNVHTLKIDS---RTVTQHSNGARNLDLKGELYIGGLSKNMFSNLP
       ::::::.: : :.  .. .:..:.   .:    ..:   :.:...:..::
gi|184 NDNQWHHVRVERNMKEA-SLQVDQLTPKTQPAPADGHVLLQLNSQLFVGG----------
          880       890        900       910       920

          1020      1030      1040      1050      1060      1070
gi|211 KLVASRD-GFQGCLASVDLNGRLPDLIADALHRIGQVERGCDGPSTTCTEESCANQGVCL
         .:.:. :: ::. :..:::   ::   : .   .:. :: :  ..  .  : : : :
gi|184 --TATRQRGFLGCIRSLQLNGMTLDLEERA-QVTPEVQPGCRGHCSSYGKL-CRNGGKCR
             930       940       950        960       970

           1080      1090      1100      1110         1120
gi|211 QQWDGFTCDCTMTSYGGPVCNDPGTTYIFGKGGALITYTWPPN---DRPST---------
       ..  :: ::::...: :: :..  ..: ::.:...: :..  :   .. :.
gi|184 ERPIGFFCDCTFSAYTGPFCSNEISAY-FGSGSSVI-YNFQENYLLSKNSSSHAASFHGD
     980       990      1000       1010       1020      1030

              1130        1140      1150       1160          1170
gi|211 -RMDRLAVGFS--THQRSAVLVRVDSASGLGDYLQLHIDQ-GTVGVIFNVGT----DDIT
```

```
        ...:  . ::  : .  ..:. :.:     .::.. : . :.. . ....     : ..
gi|184 MKLSREMIKFSFRTTRTPSLLLFVSSF--YKEYLSVIIAKNGSLQIRYKLNKYQEPDVVN
      1040      1050      1060        1070      1080      1090

           1180      1190      1200      1210      1220      1230
gi|211 IDEPNAIVSDGKYHVVRFTRSGGNATLQVDSWPVNERYPAGRQLTIFNSQAAIKIGGRDQ
       .:  :  ..::. : . ..:  : . ...:.
gi|184 FDFKN--MADGQLHHIMINREEGVVFIEIDDNRRRQVHLSSGTEFSAVKSLVLGRILEHS
        1100        1110      1120      1130      1140      1150


1642 residues in 1 query   sequences
1311 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:54:59 2002 done: Mon Dec 16 15:55:00 2002
 Scan time:  0.050 Display time:  1.917

Function used was FASTA
```

ŋ

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaKAAwxa4Wh: 1392 aa
 >gi|23498650|emb|CAC87720.2| neurexin 3-alpha [Homo sapiens]
 vs  /tmp/fastaLAAxxa4Wh library
searching /tmp/fastaLAAxxa4Wh library

   1331 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.034
The best scores are:                                       opt
gi|6694278|gb|AAF25199.1|AF193613_1 cell recognit  (1331)  429

>>gi|6694278|gb|AAF25199.1|AF193613_1 cell recognition m  (1331 aa)
 initn: 419 init1: 192 opt: 429
Smith-Waterman score: 868;  26.006% identity in 1019 aa overlap (255-1196:186-1142)

          230       240       250       260       270       280
gi|234 STTGYGGKLCSEGLSHLMMSEQGRSKAREENVATFRGSEYLCYDLSQNPIQSSSDEITLS
                                     .: .: :   : : . .. ... .: :.:.
gi|669 ARYVRIVPLDWNGEGRIGLRIEVYGCSYWADVINFDGHVVLPYRFRNKKMKTLKDVIALN
         160       170       180       190       200       210

          290       300       310       320       330
gi|234 FKTWQRNGLILH-TGKSADYVNLALKDGAVSLVINLGSGAFEAI---VEPVNGKF-NDNA
       ::: . .:.:::  :..:::..: :: . . : .::::. .  :   .  ..:.. .:.
gi|669 FKTSESEGVILHGEGQQGDYITLELKKAKLVLSLNLGSNQLGPIYGHTSVMTGSLLDDHH
         220       230       240       250       260       270

     340       350       360       370       380       390
gi|234 WHDVKVTRNLRQVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPGSPVSNNF
       ::.: . :. :......:  .     :. .. .:  :  .  :: : ..  :.:   .::
gi|669 WHSVVIERQGRSINLTLDRSMQHF-RTNGEFDYLDLDYEITFGGIPFSGK-PSSSSRKNF
         280       290        300       310       320        330

     400       410       420       430       440       450
gi|234 MGCLKEVVYKNNDIRLELSRLARIADTKMKIYGEVVFKCENVATLDPINFETPEAYISLP
        ::.. . :.. .:  .:.:  ..  ...   :.. :.: .  :. :. :..  .:. .:
gi|669 KGCMESINYNGVNIT-DLARRKKLEPSNV---GNLSFSCVEPYTV-PVFFNAT-SYLEVP
           340        350       360          370        380

     460       470       480       490       500       510
gi|234 KWNTKRMGSISFDFRTTEPNGLILFTHGKPQERKDARSQKNTKVDFFAVELLDGNLYLLL
          .. . :.::.::: .::::..:.:   .  .   .  ..::    ...  . ..  .
gi|669 GRLNQDLFSVSFQFRTWNPNGLLVFSHFADNLGNVEIDLTESKVGVH-INITQTKMSQI-
       390       400       410       420       430        440

     520       530       540       550       560       570
gi|234 DMGSGTIKVKATQKKANDGEWYHVDIQRDGRSGTISVNSRRTPFTASGESEILDLEGDMY
       :..::.          :::.:..: .      . .....  ..  . ..    .     ..
gi|669 DISSGS--------GLNDGQWHEVRFLAKENFAILTIDGDEASAVRTNSPLQVKTGEKYF
         450                 460       470       480       490

     580       590       600       610       620       630
gi|234 LGGLPENRAGLILPTELWTAMLNYGYVGCIRDLFIDGRSKNIRQLAEMQNAAGVKSSCSR
```

```
       .::.      :   ..   ..:. .. ::.. . .: .  :. ..:. . .. .. : .
gi|669 FGGF------LNQMNNSSHSVLQPSFQGCMQLIQVDDQLVNLYEVAQRKPGSFANVSID-
       500             510       520       530       540       550

     640         650       660       670          680
gi|234 MSA--KQCDSYPCKNNAVCKDGWNRFICDCTGTGYWGRTCE---REASILSY--------
       : :   .:    :.... :.. :. : : :  ::: : ::.    : :  .:
gi|669 MCAIIDRCVPNHCEHGGKCSQTWDSFKCTCDETGYSGATCHNSIYEPSCEAYKHLGQTSN
              560       570       580       590       600       610

                         690       700       710         720
gi|234 ------DGS-----------MYMKIIMPMVMHTEAEDVSFRFMSQRAYGL--LVATTSRD
             :::           :    .  .: :    ..    .. . :..  :: ..: :
gi|669 YYWIDPDGSGPLGPLKVYCNMTEDKVWTIVSHDLQMQTPVVGYNPEKYSVTQLVYSASMD
              620       630       640       650       660       670

         730       740             750        760       770
gi|234 --SADTLRLELDGGRV----KL--MVNLGKG-PETLYAGQKLNDNEWHTVRVVRRGKSLK
         :: :   :     :    :.  ..:   : : : ..:.    :: :       : ...
gi|669 QISAITDSAEYCEQYVSYFCKMSRLLNTPDGSPYTWWVGKA---NEKH-YYWGGSGPGIQ
              680       690       700       710           720

      780       790       800       810       820       830
gi|234 LTVDDDVAEGTMVGDHTRLEFHNIETGIMTEKRYISVVPSSFIGHLQSLMFNGLLYIDLC
              : :   .      . : ..     ..      ..:... . :  . ..  :
gi|669 -----KCACGIERNCTDPKYYCNCDADYKQWRK-----DAGFLSYKDHLPVSQVVVGDTD
             730       740       750            760       770

      840          850       860       870       880       890
gi|234 KNGD---IDYCELKARFGLRNIIADPVTFKTKSSYLSLATLQAYTSMHLFFQFKTTSPDG
       ..:.   ..   :. . : ::   . ..: . :::: ..:.:. ::  . : ::: .: :
gi|669 RQGSEAKLSVGPLRCQ-GDRNYW-NAASFPNPSSYLHFSTFQGETSADISFYFKTLTPWG
        780       790         800       810       820       830

         900       910        920       930       940       950
gi|234 FILFNSGDGNDFIAVELVKGY-IHYVFDLGNGPNVIKGNSDRPLNDNQWHNVVITRDNSN
        .: : :  .::: .:: ..  . . ::.::::  :   :  ::::.::: :.  : : .
gi|669 VFLENMGK-EDFIKLELKSATEVSFSFDVGNGPVEIVVRSPTPLNDDQWHRVTAER-NVK
          840        850       860       870       880        890

          960          970       980       990      1000      1010
gi|234 THSLKVD---TKVVTQVINGAKNLDLKGDLYMAGLAQGMYSNLPKLVASRDGFQGCLASV
         ::.::    ..     .:   :.: ..:...: : :           ..:: ::. :.
gi|669 QASLQVDRLPQQIRKAPTEGHTRLELYSQLFVGG-AGG-----------QQGFLGCIRSL
            900       910       920                   930       940

            1020      1030      1040      1050      1060      1070
gi|234 DLNGRLPDLINDALHRSGQIERGCEGPSTTCQEDSCANQGVCMQQWEGFTCDCSMTSYSG
        .::   :: . :   :: :  :: :  :.   . : : : :....:..::::  :.:.:
gi|669 RMNGVTLDLEERAKVTSGFIS-GCSGHCTSYGTN-CENGGKCLERYHGYSCDCSNTAYDG
              950       960        970        980       990

             1080      1090                         1100      1110
gi|234 NQCN-DPGATYIFGKSGGLILYTW--PA-----------------NDRPSTRSDRLAVGF
       . :: : ::   : . :  . :..  ::                 :..:.  ....  .:
gi|669 TFCNKDVGA---FFEEGMWLRYNFQAPATNARDSSSRVDNAPDQQNSHPDLAQEEIRFSF
     1000         1010      1020      1030      1040      1050

            1120      1130       1140      1150        1160
gi|234 STTVKDGILVRIDSAPGLGDFLQLHIEQ-GKIGVVFNIGTV--DISIKEERTPVNDGKYH
```

```
       :::    ::. :.:     ::: . ..  :.. . .:.: .    .:  ..  . .:. :
gi|669 STTKAPCILLYISSFTT--DFLAVLVKPTGSLQIRYNLGGTREPYNIDVDHRNMANGQPH
        1060      1070        1080      1090      1100      1110

     1170      1180       1190      1200      1210      1220
gi|234 VVRFTRNGGNATLQVDNWP-VNEHYPTGNTDNERFQMVKQKIPFKYNRPVEEWLQEKGRQ
        : .::.  .  :..:..: :. : :...
gi|669 SVNITRHEKTIFLKLDHYPSVSYHLPSSSDTLFNSPKSLFLGKVIETGKIDQEIHKYNTP
          1120      1130      1140      1150      1160      1170


1392 residues in 1 query   sequences
1331 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:59:10 2002 done: Mon Dec 16 15:59:11 2002
 Scan time:  0.034 Display time:  1.833

Function used was FASTA
```

N

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAC7ai6z: 1392 aa
 >gi|23498650|emb|CAC87720.2| neurexin 3-alpha [Homo sapiens]
 vs  /tmp/fastaDAAD7ai6z library
searching /tmp/fastaDAAD7ai6z library

   1154 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                       opt
gi|16306509|ref|NP_387504.1| cell recognition mol  (1154)  240

>>gi|16306509|ref|NP_387504.1| cell recognition molecule  (1154 aa)
 initn: 327 init1: 166 opt: 240
Smith-Waterman score: 616;  24.612% identity in 967 aa overlap (251-1184:178-986)

              230       240       250       260       270       280
gi|234 TCDCSTTGYGGKLCSEGLSHLMMSEQGRSKAREENVATFRGSEYLCYDLSQNPIQSSSDE
                                     : . .:. : :.  : : :...:..   :
gi|163 PPFEARFLRFLPLAWNPRGRIGMRIEVYGCAYKSEVVYFDGQSALLYRLDKKPLKPIRDV
         150       160       170       180       190       200

              290        300       310       320       330
gi|234 ITLSFKTWQRNGLILHT-GKSADYVNLALKDGAVSLVINLGSGAFEAIVEPVN---GKF-
       :.:.::. : ::..::  :. .... : :  : . .: :.. . . . ::.   :..
gi|163 ISLKFKAMQSNGILLHREGQHGNHITLELIKGKLVFFLNSGNAKLPSTIAPVTLTLGSLL
         210       220       230       240       250       260

        340       350       360       370       380       390
gi|234 NDNAWHDVKVTRNLRQVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPGSPV
       .:. ::.: .     ::...::   :    .. : ..:  .  .  :: :.    ::
gi|163 DDQHWHSVLIELLDTQVNFTVDK-HTHHFQAKGDSSYLDLNFEISFGGIPT----PGRSR
         270       280       290        300       310           320

            400       410       420       430       440       450
gi|234 S---NNFMGCLKEVVYKNNDIRLELSRLARIADTKMKIYGEVVFKCENVATLDPINFETP
       .   ..: ::::.. :.. :.    ..:.    .. ..:.: :.: .  :. :..: .
gi|163 AFRRKSFHGCLENLYYNGVDV----TELAKKHKPQILMMGNVSFSCPQPQTV-PVTFLSS
            330       340           350       360       370

            460       470       480       490       500       510
gi|234 EAYISLPKWNTKRMGSISFDFRTTEPNGLILFTHGKPQERKDARSQKNTKVDFFAVELLD
       ..:..::  . .   :..:.::: .  : .::     . :. . :        :.. : :
gi|163 RSYLALPGNSGEDKVSVTFQFRTWNRAGHLLFG----ELRRGSGS--------FVLFLKD
         380       390       400       410                   420

             520       530       540       550       560
gi|234 GNLYL-LLDMGSGTIKVKATQKKANDGEWYHVDIQRDGRSGTISVNSRRT--PFTASGES
       :.: : :.. :..  .: :     :::.:. :...      .. :..  .  :..:
gi|163 GKLKLSLFQPGQSPRNVTAGAG-LNDGQWHSVSFSAKWSHMNVVVDDDTAVQPLVA----
           430       440        450       460       470       480

     570       580       590       600       610       620
gi|234 EILDLEGDMYLGGLPENRAGLILPTELWTAMLNYGYVGCIRDLFIDGRSKNIRQLAEMQN
```

```
        .:   :: :  :              ::.. . :  .       .:. ..  ..:   .
gi|163 -VLIDSGDTYYFG-----------DAAWTVVQHGGPDAVTLRGAPSGHPRSAVSFAYAAG
               490                  500       510       520

     630       640       650       660       670        680
gi|234 AAGVKSSCSRMSAKQCDSYPCKNNAVCKDGWNRFICDCTGTGYW-GRTCEREASILSYDG
       :. ..:. .   :..:..      .. .   .:   : :  ..: ::: : ..    . :
gi|163 AGQLRSAVNL--AERCEQRLALRCGTARRPDSR---DGTPLSWWVGRTNETHTY---WGG
      530         540       550          560       570          580

      690       700       710       720       730       740
gi|234 SMYMKIIMPMVMHTEAEDVSFRFMSQRAYGLLVATTSRDSADTLRLELDGGRVKLMVNLG
       :.      :     .:.  .         ::    .  ::      . :.::     .
gi|163 SL------P-----DAQKCTC--------GL--EGNCIDSQ--YYCNCDAGR--------
                         590                 600

      750       760       770       780       790        800
gi|234 KGPETLYAGQKLNDNEWHTVRVVRRGKSLKLTVDDDVAEGTMVGD-HTRLEFHNIETGIM
                     ::: .  .:   :  .: : . :   :  :. :.. ..
gi|163 --------------NEWTSDTIVLSQKE-HLPVTQIVMTDT--GQPHSEADYT-------
                   610       620        630         640

       810       820       830       840       850       860
gi|234 TEKRYISVVPSSFIGHLQSLMFNGLLYIDLCKNGDIDYCELKARFGLRNIIADPVTFKTK
                    .: :            ::. :: ..              . ..:.:.
gi|163 -------------LGPL------------LCR-GDQSF-------------WNSASFNTE
                                  650                     660

       870       880       890       900       910        920
gi|234 SSYLSLATLQAYTSMHLFFQFKTTSPDGFILFNSGDGNDFIAVEL-VKGYIHYVFDLGNG
       .::: . ....  .  . : ::::  .: .. : :  .::: .:: .   . . ::.:::
gi|163 TSYLHFPAFHGELTADVCFFFKTTVSSGVFMENLGI-TDFIRIELRAPTEVTFSFDVGNG
        670       680       690       700        710       720

        930       940       950       960          970       980
gi|234 PNVIKGNSDRPLNDNQWHNVVITRDNSNTHSLKVDT---KVVTQVINGAKNLDLKGDLYM
       :  .  .:  :.::::::.:   : : .  ::.::    :.     .:   :.:...:..
gi|163 PCEVTVQSPTPFNDNQWHHVRAER-NVKGASLQVDQLPQKMQPAPADGHVRLQLNSQLFI
         730       740        750       760       770       780

           990      1000      1010      1020      1030      1040
gi|234 AGLAQGMYSNLPKLVASRDGFQGCLASVDLNGRLPDLINDALHRSGQIERGCEGPSTTCQ
       .: :           . . :: ::. :..:::   :: . :    : .: :: :  .:
gi|163 GGTA-----------TRQRGFLGCIRSLQLNGVALDLEERATVTPG-VEPGCAGHCSTYG
                     790       800       810        820       830

          1050      1060      1070      1080      1090      1100
gi|234 EDSCANQGVCMQQWEGFTCDCSMTSYSGNQCNDPGATYIFGKSGGLILYTWPANDRPSTR
       .  : : : : .. .: ::::....:.:  :..  ..:.   .:. . : .  .   :
gi|163 H-LCRNGGRCREKRRGVTCDCAFSAYDGPFCSNEISAYF--ATGSSMTYHFQEHYTLSEN
             840       850       860       870         880

          1110         1120        1130      1140      1150
gi|234 SDRLAVGFS---TTVKDGILV--RIDSAPGLGDFLQLHIEQGKIGVVFNIGTVDISIKEE
       :. :. ..    : ... : .  :   .:.:  ...   :.    .. : :...:  : .
gi|163 SSSLVSSLHRDVTLTREMITLSFRTTRTPSLLLYVSSFYEEYLSVILANNGSLQIRYKLD
     890       900       910       920       930       940

       1160               1170      1180      1190      1200
gi|234 R--TP---------VNDGKYHVVRFTRNGGNATLQVDNWPVNEHYPTGNTDNERFQMVKQ
```

```
       :  .:         . ::. : :...:. . . ..:..
gi|163 RHQNPDAFTFDFKNMADGQLHQVKINREEAVVMVEVNQSTKKQVILSSGTEFNAVKSLIL
     950       960       970       980       990      1000

      1210      1220      1230      1240      1250      1260
gi|234 KIPFKYNRPVEEWLQEKGRQLTIFNTQAQIAIGGKDKGRLFQGQLSGLYYDGLKVLNMAA

gi|163 GKVLEAAGADPDTRRAATSGFTGCLSAVRFGRAAPLKAALRPSGPSRVTVRGHVAPMARC
    1010      1020      1030      1040      1050      1060


1392 residues in 1 query   sequences
1154 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:58:10 2002 done: Mon Dec 16 15:58:11 2002
 Scan time:  0.050 Display time:  1.567

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaKAAgTaWrn: 1392 aa
 >gi|23498650|emb|CAC87720.2| neurexin 3-alpha [Homo sapiens]
 vs  /tmp/fastaLAAhTaWrn library
searching /tmp/fastaLAAhTaWrn library

   1311 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.066
The best scores are:                                      opt
gi|18496979|ref|NP_207837.1| cell recognition pro  (1311)  498

>>gi|18496979|ref|NP_207837.1| cell recognition protein   (1311 aa)
 initn: 303 init1: 163 opt: 498
Smith-Waterman score: 862;  25.411% identity in 1035 aa overlap (251-1209:181-1145)

              230       240       250       260       270       280
gi|234 TCDCSTTGYGGKLCSEGLSHLMMSEQGRSKAREENVATFRGSEYLCYDLSQNPIQSSSDE
                                     : . .:. . :.  : : ..:. ..  .:
gi|184 PSIKARFLRFIPLEWNPKGRIGMRIEVFGCAYRSEVVDLDGKSSLLYRFDQKSLSPIKDI
              160       170       180       190       200       210

              290       300       310       320       330
gi|234 ITLSFKTWQRNGLILHT-GKSADYVNLALKDGAVSLVINLGSGAFEAIVEPVN---GKF-
       :.:.::: : .:..::  : ..:...: :. . . :.:: : . . .    ::   :..
gi|184 ISLKFKTMQSDGILLHREGPNGDHITLQLRRARLFLLINSGEAKLPSTSTLVNLTLGSLL
              220       230       240       250       260       270

         340       350       360       370       380       390
gi|234 NDNAWHDVKVTRNLRQVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPGSPV
       .:. ::.: . :  .::...::       ... ...... :  .  :: :.    ::. :
gi|184 DDQHWHSVLIQRLGKQVNFTVDE-HRHHFHARGEFNLMNLDYEISFGGIPA----PGKSV
              280       290       300       310       320

            400       410       420       430       440       450
gi|234 S---NNFMGCLKEVVYKNNDIRLELSRLARIADTKMKIYGEVVFKCENVATLDPINFETP
       :    :: :::... :.. :: ..:..  .    ..  .:.: :.: .  .. :..: .
gi|184 SFPHRNFHGCLENLYYNGVDI-IDLAKQQK---PQIIAMGNVSFSCSQPQSM-PVTFLSS
         330       340        350          360       370        380

            460       470       480       490       500       510
gi|234 EAYISLPKWNTKRMGSISFDFRTTEPNGLILFTHGKPQERKDARSQKNTKVDFFAVELLD
       ..:..:: .. ..  : .:.::: .  ::.::.        . .  ..  . :..   :
gi|184 RSYLALPDFSGEEEVSATFQFRTWNKAGLLLFS--------ELQLISGGILLFLSDGKLK
              390       400       410               420       430

                520       530       540       550       560
gi|234 GNLY----LLLDMGSGTIKVKATQKKANDGEWYHVDIQRDGRSGTISVNSRRTPFTASGE
       .:::    :  :. .:.        . :::.:. :...      ...:... .  .
gi|184 SNLYQPGKLPSDITAGV--------ELNDGQWHSVSLSAKKNHLSVAVDGQMASAAPLLG
            440               450       460       470       480

      570       580       590       600       610       620
gi|234 SEILDLEGDMYLGGLPENRAGLILPTELWTAMLNYGYVGCIRDLFIDGRSKNI--RQLAE
```

```
        : .   : .:.:: :..  :    . :       :. ::.: . :.:.  ..   : .
gi|184 PEQIYSGGTYYFGGCPDKSFGSKCKSPLG------GFQGCMRLISISGKVVDLISVQQGS
          490       500       510             520       530

        630        640       650       660       670          680
gi|234 MQNAAGVK-SSCSRMSAKQCDSYPCKNNAVCKDGWNRFICDCTGTGYWGRTCE---REAS
       . : . .. .::. .: .   .: :.... :...:. : :.::.::: : ::.    : :
gi|184 LGNFSDLQIDSCG-ISDRCLPNY-CEHGGECSQSWSTFHCNCTNTGYRGATCHNSIYEQS
      540       550        560        570       580       590

                  690            700       710       720        730
gi|234 ILSY------DGSMYMKI-----IMPMVMHTEAEDVSFRFMSQRAYGLL-VATTSRDSAD
         .:      .: .:.       . :.... .  .... .... .  :  : .:. ..
gi|184 CEAYKHRGNTSGFYYIDSDGSGPLEPFLLYCNMTETAWTIIQHNGSDLTRVRNTNPENPY
        600       610       620       630       640       650

              740       750              760            770
gi|234 TLRLELDGGRVKLMVNLGKGP----ETLYAGQK---LNDNE-----WHTVRVVRR-----
       .  .:  ..  .:.......     :  :  .:   .: ..     : . :. .
gi|184 AGFFEYVASMEQLQATINRAEHCEQEFTYYCKKSRLVNKQDGTPLSWWVGRTNETQTYWG
        660       670       680       690       700       710

           780       790              800       810         820
gi|234 GKSLKLTVDDDVAEGTMVG-------DHTRLEFHNIETGIMTEKRYISV--VPSSFIGHL
       :.:  :       ::. .        :  : :. : .::... :... :  .  .  :.:
gi|184 GSSPDLQKCTCGLEGNCIDSQYYCNCDADRNEWTN-DTGLLAYKEHLPVTKIVITDTGRL
        720       730       740       750        760       770

          830       840       850       860       870       880
gi|234 QSLMFNGLLYIDLCKNGDIDYCELKARFGLRNIIADPVTFKTKSSYLSLATLQAYTSMHL
       .:     :  . ::. :: ..              . ..: :..::: . :...  :  .
gi|184 HSEAAYKLGPL-LCR-GDRSF-------------WNSASFDTEASYLHFPTFHGELSADV
         780         790                    800       810       820

          890       900       910        920       930       940
gi|234 FFQFKTTSPDGFILFNSGDGNDFIAVELVKG-YIHYVFDLGNGPNVIKGNSDRPLNDNQW
        : ::::. .: .: : : . ::: .:: .   . . ::.::::  :. .:   .:::::
gi|184 SFFFKTTASSGVFLENLGIA-DFIRIELRSPTVVTFSFDVGNGPFEISVQSPTHFNDNQW
              830       840        850       860       870

           950       960          970       980       990      1000
gi|234 HNVVITRDNSNTHSLKVD---TKVVTQVINGAKNLDLKGDLYMAGLAQGMYSNLPKLVAS
       :.: . : : .  ::.::    :.     .:   :.:...:...: :           .
gi|184 HHVRVER-NMKEASLQVDQLTPKTQPAPADGHVLLQLNSQLFVGGTA-----------TR
     880        890       900       910       920

             1010      1020      1030      1040      1050      1060
gi|234 RDGFQGCLASVDLNGRLPDLINDALHRSGQIERGCEGPSTTCQEDSCANQGVCMQQWEGF
       . :: ::. :..:::   :: . : . . ... ::.:  ..  .  : : : : ..  ::
gi|184 QRGFLGCIRSLQLNGMTLDLEERA-QVTPEVQPGCRGHCSSYGK-LCRNGGKCRERPIGF
       930       940       950        960       970        980

             1070      1080      1090      1100      1110
gi|234 TCDCSMTSYSGNQCNDPGATYIFGKSGGLILYTWPANDRPSTRSDRLAVGFSTTVKDG--
        :::....:.:  :..  ..: :: ::. ..:..  :   :  :.  :..:   .: .
gi|184 FCDCTFSAYTGPFCSNEISAY-FG-SGSSVIYNFQENYLLSKNSSSHAASFHGDMKLSRE
         990      1000        1010      1020      1030      1040

        1120      1130      1140      1150          1160
gi|234 ---ILVRIDSAPGLGDFLQLHIEQGKIGVVFNIGTVDISIK----EERTPVN-------D
```

```
          .  :   .:.:  :..   ..    .. . :...:  :    .:   ::       :
gi|184 MIKFSFRTTRTPSLLLFVSSFYKEYLSVIIAKNGSLQIRYKLNKYQEPDVVNFDFKNMAD
          1050      1060      1070      1080      1090      1100

         1170      1180      1190      1200      1210      1220
gi|234 GKYHVVRFTRNGGNATLQVDNWPVNEHYPTGNTDNERFQMVKQKIPFKYNRPVEEWLQEK
       :. : . ..:. : . ...:.   :..  .  ... .:. ::. .
gi|184 GQLHHIMINREEGVVFIEIDD---NRRRQVHLSSGTEFSAVKSLVLGRILEHSDVDQETA
          1110      1120         1130      1140      1150      1160

         1230      1240      1250      1260      1270      1280
gi|234 GRQLTIFNTQAQIAIGGKDKGRLFQGQLSGLYYDGLKVLNMAAENNPNIKINGSVRLVGE

gi|184 LAGAQGFTGCLSAVQLSHVAPLKAALHPSHPDPVTVTGHVTESSCMAQPGTDATSRERTH
             1170      1180      1190      1200      1210      1220


1392 residues in 1 query   sequences
1311 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:57:28 2002 done: Mon Dec 16 15:57:29 2002
 Scan time:  0.066 Display time:  1.817

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAmhaaYv: 1345 aa
 >FIRST_SEQUENCE
 vs  /tmp/fastaDAAnhaaYv library
searching /tmp/fastaDAAnhaaYv library

   1331 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:. 0.050
The best scores are:                                       opt
gi|6694278|gb|AAF25199.1|AF193613_1 cell recognit  (1331) 4412

>>gi|6694278|gb|AAF25199.1|AF193613_1 cell recognition m  (1331 aa)
 initn: 3862 init1: 2022 opt: 4412
Smith-Waterman score: 4577;  50.724% identity in 1313 aa overlap (66-1345:30-1331)

          40        50        60        70        80        90
FIRST_ INLIKEMDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHS
                                     .:. .::.::.: :  .::::::...:..:
gi|669  MQAAPRAGCGAALLLWIVSSCLCRAWTAPSTSQKCDEPLVSGLPHVAFSSSSSISGSYS
                10        20        30        40        50

           100       110       120       130       140       150
FIRST_ P--AQLNWRVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSD
       :  :..: : :.:::::.::.  ::::.:.::: .:.:.::::::.::::::.: ...::
gi|669 PGYAKINKRGGAGGWSPSDSDHYQWLQVDFGNRKQISAIATQGRYSSSDWVTQYRMLYSD
      60        70        80        90       100       110

           160       170       180       190       200       210
FIRST_ TGRNWKQYKQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVY
       :::::: :.:. .::.: ::.:.:.::.:.: : . ::.::.:::.::  :.::.:.:::
gi|669 TGRNWKPYHQDGNIWAFPGNINSDGVVRHELQHPIIARYVRIVPLDWNGEGRIGLRIEVY
     120       130       140       150       160       170

           220       230       240       250       260       270
FIRST_ GCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLE
       :::: .:: .:::.  : ::: .: :.::::::.:.::. ...::..:::::.::.::::
gi|669 GCSYWADVINFDGHVVLPYRFRNKKMKTLKDVIALNFKTSESEGVILHGEGQQGDYITLE
     180       190       200       210       220       230

           280       290       300       310       320       330
FIRST_ LQKGRLALHLNLGDSKARLSSSLPSATLGSLLDDQHWH-VLIERVGKQVNFTVDKHTQHF
       :.:..:.: ::::...     .  :.  ::::::.::: :.::: :...:.:.:.  :::
gi|669 LKKAKLVLSLNLGSNQLGPIYGHTSVMTGSLLDDHHWHSVVIERQGRSINLTLDRSMQHF
     240       250       260       270       280       290

            340       350       360       370       380       390
FIRST_ RTKGETDALDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNII-LAKRRKHQIY
       ::.:: : ::.:::..:::::  :::..  .:::.::.:.. ::::::  ::.:.: .
gi|669 RTNGEFDYLDLDYEITFGGIPFSGKPSSSSRKNFKGCMESINYNGVNITDLARRKKLEPS
     300       310       320       330       340       350

             400       410       420       430       440       450
FIRST_ TVGNVTFSCSEPQIVPITFNSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEG
```

```
       .::::..:: ::  ::. ::.. ::: .::  . : .:::::::::: .:::. ......
gi|669 NVGNLSFSCVEPYTVPVFFNAT-SYLEVPGRLNQDLFSVSFQFRTWNPNGLLVFSHFADN
     360       370       380        390       400       410

                460        470       480       490       500
FIRST_ SGTLLLSL---EGGI-LRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDD
        :.. ..:   . :. . ..  ::..   .: .::.:::: :: : . :..:   ::.:
gi|669 LGNVEIDLTESKVGVHINITQTKMSQ--IDISSGSGLNDGQWHEVRFLAKENFAILTIDG
      420       430       440         450       460       470

       510       520       530       540         550       560
FIRST_ EAAPPAPDSTWVQIYSGNSYYFGGCPDNLTDSQ--CLNPIKAFQGCMRLIFIDNQPKDLI
       . :  .  .. .:. .:..:.:::  .....:.   :.:  .:::::.:: .:.:  .:
gi|669 DEASAVRTNSPLQVKTGEKYFFGGFLNQMNNSSHSVLQP--SFQGCMQLIQVDDQLVNLY
        480       490       500       510         520       530

         570       580       590       600       610       620
FIRST_ SVQQGSLGNFSDLHIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSI
        : : . :.:... ::.:.: :::.::.:::::.:::.: .: :.:..:.:.::::::::
gi|669 EVAQRKPGSFANVSIDMCAIIDRCVPNHCEHGGKCSQTWDSFKCTCDETGYSGATCHNSI
          540       550       560       570       580       590

         630       640       650       660       670       680
FIRST_ YEQSCEVYRHQGNTAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGAN
       :: :::.:.: :.:.....:: :::::::::.::::.::::.:: :.:.    : : : :
gi|669 YEPSCEAYKHLGQTSNYYWIDPDGSGPLGPLKVYCNMTEDKVWTIVSHDLQMQTPVVGYN
          600       610       620       630       640       650

         690       700       710       720       730       740
FIRST_ PEKPYAMALDYGGSMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNER
       :::  .  : :..::.:. :. :..:.::: :.: :. :::::::::.:.:::.:..::.
gi|669 PEKYSVTQLVYSASMDQISAITDSAEYCEQYVSYFCKMSRLLNTPDGSPYTWWVGKANEK
          660       670       680       690       700       710

         750       760       770       780       790       800
FIRST_ HPYWGGSPPGVQQCECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITD
       : ::::: ::.:.: ::....: :  ..::::::  .: .:.::::.::::::.:.:.:. :
gi|669 HYYWGGSGPGIQKCACGIERNCTDPKYYCNCDADYKQWRKDAGFLSYKDHLPVSQVVVGD
          720       730       740       750       760       770

         810       820       830       840       850       860
FIRST_ TDRSNSEAAWRIGPLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSG
       :::..:::   .::::: ::: .:::.:: . .::::: ::..: ::::::.::: .  :
gi|669 TDRQGSEAKLSVGPLRCQGDRNYWNAASFPNPSSYLHFSTFQGETSADISFYFKTLTPWG
          780       790       800       810       820       830

         870       880       890       900       910       920
FIRST_ VFLENLGIKDFIRLEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKET
       :::::.: .:::.::..: .:..:..::::::::.::.::. :::.::: : ::::.:..
gi|669 VFLENMGKEDFIKLELKSATEVSFSFDVGNGPVEIVVRSPTPLNDDQWHRVTAERNVKQA
          840       850       860       870       880       890

         930       940       950       960       970       980
FIRST_ SLQVDNLPRSTRETSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERA
       ::::: ::.. :..  ::: ::.: :::::::... :.:::::::::..::  .::::::
gi|669 SLQVDRLPQQIRKAPTEGHTRLELYSQLFVGGAGG-QQGFLGCIRSLRMNGVTLDLEERA
          900       910       920       930       940       950

         990      1000      1010      1020      1030      1040
FIRST_ KVTSGVRPGCPGHCSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGT
```

```
       :::::   :: :::.:::. :.:::::.:...:: :::.:. :.: ::.:.:.: :: :
gi|669 KVTSGFISGCSGHCTSYGTNCENGGKCLERYHGYSCDCSNTAYDGTFCNKDVGAFFEEGM
           960       970       980       990      1000      1010

        1050      1060      1070               1080      1090
FIRST_ SVTYMFQEPYPVTKNISLSSSAIYTDSAPSKEN---------IALSFVTTQAPSLLLFIN
        . : :: :   . :   ::: .  :.::...:       . : .:: ::.:: .::.:.
gi|669 WLRYNFQAP---ATNARDSSSRV--DNAPDQQNSHPDLAQEEIRFSFSTTKAPCILLYIS
          1020         1030        1040      1050      1060

       1100      1110      1120       1130      1140      1150
FIRST_ SSSQDFVVVLLCKNGSLQVRYHLN-KEETHVFTIDADNFANRRMHHLKINREGRELTIQM
       : . ::..::.  .::::.::.:.  .: . . .:  :.:: . : ..:.:. . . ...
gi|669 SFTTDFLAVLVKPTGSLQIRYNLGGTREPYNIDVDHRNMANGQPHSVNITRHEKTIFLKL
      1070      1080      1090      1100      1110      1120

        1160        1170      1180      1190      1200      1210
FIRST_ DQQLRLSYNF--SPEVEFRVIRSLTLGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIA
       :.   .::..  : .. :   .:: :::: :.  .:.:. : :. ::.::.: ::.:.::
gi|669 DHYPSVSYHLPSSSDTLFNSPKSLFLGKVIETGKIDQEIHKYNTPGFTGCLSRVQFNQIA
      1130      1140      1150      1160      1170      1180

          1220       1230      1240        1250            1260
FIRST_ PLKAALRHATV-APVTVHGTLTESSCGF--MVDSDVNAVTT------VHSSSDPFG-KTD
       :::::::.... : : ..: :.::.::   .. : ....:       . :.:  :  .
gi|669 PLKAALRQTNASAHVHIQGELVESNCGASPLTLSPMSSATDPWHLDHLDSASADFPYNPG
      1190      1200      1210      1220      1230      1240

          1270      1280      1290      1300      1310      1320
FIRST_ EREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMTRFLYQHKQSHRTSQMKEKEYPENLD
       . . . :.:  .::.:::::::::: :.: . .. :....:: ...:.. :  :  :. :
gi|669 QGQAIRNGVNRNSAIIGGVIAVVIFTILCTLVFLIRYMFRHKGTYHTNEAKGAESAESAD
      1250      1260      1270      1280      1290      1300

           1330      1340
FIRST_ SSF-RNEIDLQNTVSECKREYFI
       ...  :. .. .:..: :.:..:
gi|669 AAIMNNDPNFTETIDESKKEWLI
      1310      1320      1330


1345 residues in 1 query   sequences
1331 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:40:18 2002 done: Mon Dec 16 15:40:19 2002
 Scan time:  0.050 Display time:  2.484

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAZ8aisO: 1345 aa
 >FIRST_SEQUENCE
 vs  /tmp/fastaHAA08aisO library
searching /tmp/fastaHAA08aisO library

   1154 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.050
The best scores are:                                         opt
gi|16306509|ref|NP_387504.1| cell recognition mol   (1154) 2301

>>gi|16306509|ref|NP_387504.1| cell recognition molecule  (1154 aa)
 initn: 3683 init1: 1805 opt: 2301
Smith-Waterman score: 3850;  47.578% identity in 1280 aa overlap (51-1317:11-1152)

               30        40        50        60        70        80
FIRST_ PTPBETAANDNEREXINLIKEMDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLS
                                     :: :: .  :    ...  .:: :::: :
gi|163                     MASVAWAVLKVLLLLPTQTWSPVGAGNPPDCDAPLASALP
                                     10        20        30        40

               90         100       110       120       130
FIRST_ PMAFSSSSDLTGTHSP--AQLNWRVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRY
         .:::::.:...:.:  ..:: : :.:::.:  ::  ::::.:::.:.:.::::::: :
gi|163 RSSFSSSSELSSSHGPGFSRLNRRDGAGGWTPLVSNKYQWLQIDLGERMEVTAVATQGGY
               50        60        70        80        90       100

      140       150       160       170       180       190
FIRST_ GSSDWVTSYSLMFSDTGRNWKQYKQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPL
       ::::::::: ::::: :::::::..:.::: : :: :::::::..:    .:::.::.::
gi|163 GSSDWVTSYLLMFSDGGRNWKQYRREESIWGFPGNTNADSVVHYRLQPPFEARFLRFLPL
              110       120       130       140       150       160

      200       210       220       230       240       250
FIRST_ EWNPSGKIGMRVEVYGCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGV
        ::: :.::::.:::::.:::.:. :::.:.::::...: .. ..::::::::.::..:.
gi|163 AWNPRGRIGMRIEVYGCAYKSEVVYFDGQSALLYRLDKKPLKPIRDVISLKFKAMQSNGI
              170       180       190       200       210       220

      260       270       280       290         300       310
FIRST_ LFHGEGQRGDHITLELQKGRLALHLNLGDSKARLSSSL-P-SATLGSLLDDQHWH-VLIE
       :.: :::.:.:::::: ::.:.. :: :.  :.: :.. : . :::::::::::: ::::
gi|163 LLHREGQHGNHITLELIKGKLVFFLNSGN--AKLPSTIAPVTLTLGSLLDDQHWHSVLIE
              230       240         250       260       270

          320       330       340       350       360       370
FIRST_ RVGKQVNFTVDKHTQHFRTKGETDALDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYY
        .  ::::::::::.::..::... ::...:.::::::.::.  .: .:.::::.:::::
gi|163 LLDTQVNFTVDKHTHHFQAKGDSSYLDLNFEISFGGIPTPGRSRAFRRKSFHGCLENLYY
      280       290       300       310       320       330

          380        390       400       410       420       430
FIRST_ NGVNII-LAKRRKHQIYTVGNVTFSCSEPQIVPITFNSSGSYLLLPGTPQIDGLSVSFQF
```

```
       :::..  :::..: ::  .:::.::: .:: ::.:: :: ::: :::.   : .::.:::
gi|163 NGVDVTELAKKHKPQILMMGNVSFSCPQPQTVPVTFLSSRSYLALPGNSGEDKVSVTFQF
      340       350       360       370       380       390

          440       450       460       470       480       490
FIRST_ RTWNKDGLLLSTELSEGSGTLLLSLEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSI
       ::::. : ::  :: .:::...: :. : :.: . .  .   .. .:..:::: :::::.
gi|163 RTWNRAGHLLFGELRRGSGSFVLFLKDGKLKLSLFQPGQSPRNVTAGAGLNDGQWHSVSF
      400       410       420       430       440       450

          500       510       520       530       540       550
FIRST_ NARRNRITLTLDDEAAPPAPDSTWVQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRL
       .:. .......::..:  .   . : : ::..::::
gi|163 SAKWSHMNVVVDDDTA--VQPLVAVLIDSGDTYYFG------------------------
      460       470         480       490

          560       570       580       590       600       610
FIRST_ IFIDNQPKDLISVQQGSLGNFSDLHIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDT

gi|163 ------------------------------------------------------------


          620       630       640       650       660       670
FIRST_ SYTGATCHNSIYEQSCEVYRHQGNTAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHN
                                                         :  :: :::.
gi|163 --------------------------------------------------DAAWTVVQHG
                                                              500

          680       690        700       710       720       730
FIRST_ NTELTRVRGANPEKPY-AMALDYGGSMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGT
       . . . .:::   .:  :... :...  ::..... .:.:::..: .:  .:  .. :::
gi|163 GPDAVTLRGAPSGHPRSAVSFAYAAGAGQLRSAVNLAERCEQRLALRCGTARRPDSRDGT
            510       520       530       540       550       560

           740       750       760       770       780       790
FIRST_ PFTWWIGRSNERHPYWGGSPPGVQQCECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFK
       :..::.::.:: : ::::: : .:.: :::. .:.: :..::::: ..:::.::  :: :
gi|163 PLSWWVGRTNETHTYWGGSLPDAQKCTCGLEGNCIDSQYYCNCDAGRNEWTSDTIVLSQK
            570       580       590       600       610       620

           800       810       820       830       840       850
FIRST_ DHLPVTQIVITDTDRSNSEAAWRIGPLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSAD
       .::::::::.::: . .::: . .::: : ::. :::..:: ::.::::::.::.:..::
gi|163 EHLPVTQIVMTDTGQPHSEADYTLGPLLCRGDQSFWNSASFNTETSYLHFPAFHGELTAD
            630       640       650       660       670       680

           860       870       880       890       900       910
FIRST_ ISFFFKTTALSGVFLENLGIKDFIRLEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQW
       . ::::::. ::::.::::: ::::.:. .:.:.::. :::::: :..::::. .:::::
gi|163 VCFFFKTTVSSGVFMENLGITDFIRIELRAPTEVTFSFDVGNGPCEVTVQSPTPFNDNQW
            690       700       710       720       730       740

           920       930       940       950       960       970
FIRST_ HYVRAERNLKETSLQVDNLPRSTRETSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLH
       :.::::::.: .:::::.::.. . .  .:: :::::::::.:::..::.:::::::::.
gi|163 HHVRAERNVKGASLQVDQLPQKMQPAPADGHVRLQLNSQLFIGGTATRQRGFLGCIRSLQ
            750       760       770       780       790       800

           980       990      1000      1010      1020      1030
FIRST_ LNGQKMDLEERAKVTSGVRPGCPGHCSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFC
```

```
       :::  .:::::: :: ::.::: :::::.:: .:.:::.: ::. :  :::. : :.::::
gi|163 LNGVALDLEERATVTPGVEPGCAGHCSTYGHLCRNGGRCREKRRGVTCDCAFSAYDGPFC
               810       820       830       840       850       860

             1040      1050      1060      1070      1080      1090
FIRST_ KKEVSAVFEAGTSVTYMFQEPYPVTKNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLL
       ..:.:: : .:.:.:: ::: : ...: :   :... : . ..: :.::: ::..:::::
gi|163 SNEISAYFATGSSMTYHFQEHYTLSENSSSLVSSLHRDVTLTREMITLSFRTTRTPSLLL
               870       880       890       900       910       920

             1100      1110      1120       1130      1140      1150
FIRST_ FINSSSQDFVVVLLCKNGSLQVRYHLNKEET-HVFTIDADNFANRRMHHLKINREGRELT
       ...:  .... :.: .::::::.::.:....  .::.:  :.:. ..:..:::::   .
gi|163 YVSSFYEEYLSVILANNGSLQIRYKLDRHQNPDAFTFDFKNMADGQLHQVKINREEAVVM
               930       940       950       960       970       980

              1160      1170      1180      1190      1200      1210
FIRST_ IQMDQQLRLSYNFSPEVEFRVIRSLTLGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHI
       ....:. . .  .:  .:: ...:: :::: :  : : .. .: . ::.::.:.:....
gi|163 VEVNQSTKKQVILSSGTEFNAVKSLILGKVLEAAGADPDTRRAATSGFTGCLSAVRFGRA
               990      1000      1010      1020      1030      1040

              1220      1230       1240         1250      1260
FIRST_ APLKAALRHATVAPVTVHGTLTE-SSCGFMVDSDVNA---VTTVHSSSDPFGKTDEREPL
       :::::::: .  . :::.: ..  . :.  . :   :   .  . ...   : .:: :::
gi|163 APLKAALRPSGPSRVTVRGHVAPMARCAAGAASGSPARELAPRLAGGAGRSGPADEGEPL
              1050      1060      1070      1080      1090      1100

        1270      1280      1290      1300       1310      1320
FIRST_ TNAVRSDSAVIGGVIAVVIFIIFCIIGIMTRFLYQHK-QSHRTSQMKEKEYPENLDSSFR
       .:: : :::::::::::::::..:: .:  :.  :.: ...  :....::
gi|163 VNADRRDSAVIGGVIAVVIFILLCITAIAIRIYQQRKLRKENESKVSKKEEC
              1110      1120      1130      1140      1150

         1330      1340
FIRST_ NEIDLQNTVSECKREYFI


1345 residues in 1 query   sequences
1154 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:29:32 2002 done: Mon Dec 16 15:29:34 2002
 Scan time:  0.050 Display time:  2.167

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAV8aisO: 1345 aa
 >FIRST_SEQUENCE
 vs  /tmp/fastaDAAW8aisO library
searching /tmp/fastaDAAW8aisO library

   1311 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.066
The best scores are:                                      opt
gi|18496979|ref|NP_207837.1| cell recognition pro  (1311) 5338

>>gi|18496979|ref|NP_207837.1| cell recognition protein   (1311 aa)
 initn: 2330 init1: 2086 opt: 5338
Smith-Waterman score: 5338;  59.062% identity in 1280 aa overlap (70-1344:33-1310)

      40        50        60        70        80        90
FIRST_ KEMDSLPRLTSVLTLLFSGLWHLGLTATNYNCDDPLASLLSPMAFSSSSDLTGTHSP--A
                                     .:::::.: :   .:::::.:...:.:  :
gi|184 LFYLLVVLSIDSTKASALTNPNVALFLLADDCDDPLVSALPQASFSSSSELSSSHGPGFA
             10        20        30        40        50        60

       100       110       120       130       140       150
FIRST_ QLNWRVGTGGWSPADSNAQQWLQMDLGNRVEITAVATQGRYGSSDWVTSYSLMFSDTGRN
       .:: : :.:::::  ::  ::::.:::.:.:.::::::: ::::.::::: :::::.: :
gi|184 RLNRRDGAGGWSPLVSNKYQWLQIDLGERMEVTAVATQGGYGSSNWVTSYLLMFSDSGWN
             70        80        90       100       110       120

       160       170       180       190       200       210
FIRST_ WKQYKQEDSIWTFAGNMNADSVVHHKLLHSVRARFVRFVPLEWNPSGKIGMRVEVYGCSY
       ::::.:::::: :.:: ::::::...:  :..:::.::.::::::.:.::::.::.::.:
gi|184 WKQYRQEDSIWGFSGNANADSVVYYRLQPSIKARFLRFIPLEWNPKGRIGMRIEVFGCAY
            130       140       150       160       170       180

       220       230       240       250       260       270
FIRST_ KSDVADFDGRSSLLYRFNQKLMSTLKDVISLKFKSMQGDGVLFHGEGQRGDHITLELQKG
       .:.:.:.::.:::::::.:: .: .::.::::::.::.::.:.: ::  ::::::.:...
gi|184 RSEVVDLDGKSSLLYRFDQKSLSPIKDIISLKFKTMQSDGILLHREGPNGDHITLQLRRA
            190       200       210       220       230       240

       280       290       300       310       320       330
FIRST_ RLALHLNLGDSKARLSSSLPSATLGSLLDDQHWH-VLIERVGKQVNFTVDKHTQHFRTKG
       :: : .: :..:   .:.: . :::::::::::: :::.:.:::::::::.: .::...:
gi|184 RLFLLINSGEAKLPSTSTLVNLTLGSLLDDQHWHSVLIQRLGKQVNFTVDEHRHHFHARG
            250       260       270       280       290       300

        340       350       360       370       380       390
FIRST_ ETDALDIDYELSFGGIPVPGKPGTFLKKNFHGCIENLYYNGVNII-LAKRRKHQIYTVGN
       : . ...:::.::::::.:::  .: ..:::::.::::::::.:: :::..: :: ..::
gi|184 EFNLMNLDYEISFGGIPAPGKSVSFPHRNFHGCLENLYYNGVDIIDLAKQQKPQIIAMGN
            310       320       330       340       350       360

         400       410       420       430       440       450
FIRST_ VTFSCSEPQIVPITFNSSGSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSEGSGTL
```

```
       :.:::::.:: .:.:: :: ::: ::    . .:..:::::::: :::: .::.  :: .
gi|184 VSFSCSQPQSMPVTFLSSRSYLALPDFSGEEEVSATFQFRTWNKAGLLLFSELQLISGGI
            370       380       390       400       410       420

         460       470       480       490       500       510
FIRST_ LLSLEGGILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAPPAPD
       :: :  : :.  . .  .  ..: .: .:::: :::::..:..:......: . :  ::
gi|184 LLFLSDGKLKSNLYQPGKLPSDITAGVELNDGQWHSVSLSAKKNHLSVAVDGQMASAAPL
            430       440       450       460       470       480

         520       530       540       550       560       570
FIRST_ STWVQIYSGNSYYFGGCPDNLTDSQCLNPIKAFQGCMRLIFIDNQPKDLISVQQGSLGNF
           :::::..::::::::.   :.: .:. .:::::::: :...  :::::::::::::
gi|184 LGPEQIYSGGTYYFGGCPDKSFGSKCKSPLGGFQGCMRLISISGKVVDLISVQQGSLGNF
            490       500       510       520       530       540

         580       590       600       610       620       630
FIRST_ SDLHIDLCSIKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRH
       :::.:: :.:.:::::::::::: :::::.::.:::..:.: ::::::::::::::.:.:
gi|184 SDLQIDSCGISDRCLPNYCEHGGECSQSWSTFHCNCTNTGYRGATCHNSIYEQSCEAYKH
            550       560       570       580       590       600

         640       650       660       670       680       690
FIRST_ QGNTAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAMALD
       .:::.::.:::::::::: :. .:::.::   :: .:::...:::::..:::.:::  ..
gi|184 RGNTSGFYYIDSDGSGPLEPFLLYCNMTETA-WTIIQHNGSDLTRVRNTNPENPYAGFFE
            610       620       630       640       650       660

         700       710       720       730       740       750
FIRST_ YGGSMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWGGSPPG
       : .:::::.:.:. .:::::: .:.:..:::.:  ::::..::.::.:: . ::::: :
gi|184 YVASMEQLQATINRAEHCEQEFTYYCKKSRLVNKQDGTPLSWWVGRTNETQTYWGGSSPD
            670       680       690       700       710       720

         760       770       780       790       800       810
FIRST_ VQQCECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPVTQIVITDTDRSNSEAAW
       .:.: :::. .:.: :..::::::..:::::::.:..:.:::::.:::::: : .::::.
gi|184 LQKCTCGLEGNCIDSQYYCNCDADRNEWTNDTGLLAYKEHLPVTKIVITDTGRLHSEAAY
            730       740       750       760       770       780

         820       830       840       850       860       870
FIRST_ RIGPLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFFKTTALSGVFLENLGIKD
       ..::: : ::: :::..:: ::::::::::::.:.:::.:::::::: :::::::::: :
gi|184 KLGPLLCRGDRSFWNSASFDTEASYLHFPTFHGELSADVSFFFKTTASSGVFLENLGIAD
            790       800       810       820       830       840

         880       890       900       910       920       930
FIRST_ FIRLEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVRAERNLKETSLQVDNLPRS
       :::.:. ::. .::..:::::: :. ::::. .::::::.::.:::.::.:::::.:  .
gi|184 FIRIELRSPTVVTFSFDVGNGPFEISVQSPTHFNDNQWHHVRVERNMKEASLQVDQLTPK
            850       860       870       880       890       900

         940       950       960       970       980       990
FIRST_ TRETSEEGHFRLQLNSQLFVGGTSSRQKGFLGCIRSLHLNGQKMDLEERAKVTSGVRPGC
       :. .  .::  ::::::::::::..::.:::::::::.:::. .::::::.::  :.:::
gi|184 TQPAPADGHVLLQLNSQLFVGGTATRQRGFLGCIRSLQLNGMTLDLEERAQVTPEVQPGC
            910       920       930       940       950       960

       1000      1010      1020      1030      1040      1050
FIRST_ PGHCSSYGSICHNGGKCVEKHNGYLCDCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPY
```

```
        ::::::..:.:::::: :.  :..:::: : : ::::..:.:: : .:.:: : ::: :
gi|184 RGHCSSYGKLCRNGGKCRERPIGFFCDCTFSAYTGPFCSNEISAYFGSGSSVIYNFQENY
             970       980       990      1000      1010      1020

        1060      1070      1080      1090      1100      1110
FIRST_ PVTKNISLSSSAIYTDSAPSKENIALSFVTTQAPSLLLFINSSSQDFVVVLLCKNGSLQV
        ..:: :  ..... :   :.: : .:: ::..::::::..:  .... :.. ::::::.
gi|184 LLSKNSSSHAASFHGDMKLSREMIKFSFRTTRTPSLLLFVSSFYKEYLSVIIAKNGSLQI
            1030      1040      1050      1060      1070      1080

        1120      1130      1140      1150      1160      1170
FIRST_ RYHLNK-EETHVFTIDADNFANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVI
       ::.::: .:  : ..:  :.:. ..::. ::::   . :..:.. : . ..:  .:: ..
gi|184 RYKLNKYQEPDVVNFDFKNMADGQLHHIMINREEGVVFIEIDDNRRRQVHLSSGTEFSAV
            1090      1100      1110      1120      1130      1140

         1180      1190      1200      1210      1220      1230
FIRST_ RSLTLGKVTENLGLDSEVAKANAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLT
       .::.::.. :.  .:.:.: :.:.::.::.:.:: .:.:::::::. .   :::: : .:
gi|184 KSLVLGRILEHSDVDQETALAGAQGFTGCLSAVQLSHVAPLKAALHPSHPDPVTVTGHVT
            1150      1160      1170      1180      1190      1200

         1240      1250      1260      1270      1280      1290
FIRST_ ESSCGFMVDSDVNAVTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCII
       ::::  .  .:...   .:: .:  :  :.::::.::..::::::::.::::::..::
gi|184 ESSCMAQPGTDATSRERTHSFADHSGTIDDREPLANAIKSDSAVIGGLIAVVIFILLCIT
            1210      1220      1230      1240      1250      1260

         1300      1310      1320      1330      1340
FIRST_ GIMTRFLYQHKQSHRTSQMKEKEYPENLDSSFRNEIDLQNTVSECKREYFI
       .: .: .::.:. .. :. :..:  .. .. ...:...::.:.: ..:::.
gi|184 AIAVR-IYQQKRLYKRSEAKRSENVDSAEAVLKSELNIQNAVNENQKEYFF
             1270      1280      1290      1300      1310


1345 residues in 1 query   sequences
1311 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:27:28 2002 done: Mon Dec 16 15:27:30 2002
 Scan time:  0.066 Display time:  2.417

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaGAAf3ayIx: 1345 aa
 >FIRST_SEQUENCE
 vs  /tmp/fastaHAAg3ayIx library
searching /tmp/fastaHAAg3ayIx library

   1477 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.067
The best scores are:                                      opt
gi|14149613|ref|NP_004792.1| neurexin 1 isoform a  (1477)  459

>>gi|14149613|ref|NP_004792.1| neurexin 1 isoform alpha   (1477 aa)
 initn: 261 init1: 142 opt: 459
Smith-Waterman score: 861;  24.780% identity in 1138 aa overlap (221-1269:283-1345)

              200       210       220       230       240       250
FIRST_ RFVRFVPLEWNPSGKIGMRVEVYGCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLKF
                                     .: : :   . : ..:. ... .: :.:.:
gi|141 KDCSQEDNNVEGLAHLMMGDQGKSKGKEEYIATFKGSEYFCYDLSQNPIQSSSDEITLSF
            260       270       280       290       300       310

              260       270       280       290       300       310
FIRST_ KSMQGDGVLFHGEGQRGDHITLELQKGRLALHLNLGDSKARLSSSLPSATLGSLLDDQHW
       :..: .:...:  :. .:...: :..: ..: .::: : :  .   :   ... ..:. :
gi|141 KTLQRNGLMLH-TGKSADYVNLALKNGAVSLVINLG-SGAFEALVEP---VNGKFNDNAW
            320        330       340        350          360

               320                330       340       350
FIRST_ H-VLIERVGKQ--------VNFTVDK-HTQHFRTKGETDALDIDYELSFGGIPVPGK-PG
       : : . :  .:        :...::   :    :. .   :  :  .  :: :  .  ::
gi|141 HDVKVTRNLRQHSGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPG
       370       380       390       400       410       420

     360       370       380            390       400        410
FIRST_ TFLKKNFHGCIENLYYNGVNIIL-----AKRRKHQIYTVGNVTFSCSE-PQIVPITFNSS
       . ...:: ::.... :.. .. :     ::.   ..   : :.:.: .   . ::::..
gi|141 SPVSNNFMGCLKEVVYKNNDVRLELSRLAKQGDPKMKIHGVVAFKCENVATLDPITFETP
       430       440       450       460       470       480

           420       430       440            450               460
FIRST_ GSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTE-----LSEGSGTLLLS--------LE
        :.. ::        :.::.::: . .::.: ..      ....   ...        :.
gi|141 ESFISLPKWNAKKTGSISFDFRTTEPNGLILFSHGKPRHQKDAKHPQMIKVDFFAIEMLD
       490       500       510       520       530       540

               470       480       490       500       510
FIRST_ GGI-LRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAP-PAPDSTW
       : . : : . . : ..  .:  ...::: :. :...      :....   .:  ::  .
gi|141 GHLYLLLDMGSGTIKIKALL--KKVNDGEWYHVDFQRDGRSGTISVNTLRTPYTAPGESE
       550       560         570       580       590       600

      520       530           540         550       560       570
FIRST_ VQIYSGNSYYFGGCPDN----LTDSQCLNPI--KAFQGCMRLIFIDNQPKDLISVQQGSL
```

```
       . .   .  :.:: :.:   .  .. .  .  .. ::.: .:::.: ::.  :.. .
gi|141 I-LDLDDELYLGGLPENKAGLVFPTEVWTALLNYGYVGCIRDLFIDGQSKDI--RQMAEV
          610       620       630       640       650         660

            580         590       600       610       620       630
FIRST_ GNFSDLHIDLCSIKDR--CLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSC
        . . .. . :: .   :: : :...: : ..:. . :.:: :.: : .:   :.
gi|141 QSTAGVKPS-CSKETAKPCLSNPCKNNGMCRDGWNRYVCDCSGTGYLGRSC-----EREA
            670        680       690       700       710

              640       650       660       670       680       690
FIRST_ EVYRHQGNTAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPY
        :  ..:.   :. :.       :. .. .  . ..   :.    :  . . .
gi|141 TVLSYDGSM--FMKIQL-------PVVMHTEAEDVSLRFRSQRAYGILMATTSRDSADTL
        720         730              740       750       760

              700       710       720       730       740       750
FIRST_ AMALDYGGSMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGRSNERHPYWG
        . :: :    .: . .:    :   .  .:  :.    :. : :. .   .:: :
gi|141 RLELDAGRV--KLTVNLD----C---IRINCNSSK---GPE-TLFAGYNLNDNEWHTVRV
       770         780              790           800       810

              760       770       780        790          800
FIRST_ GSPPGVQQCECGLDESCLDIQHFCNCDADKDEWTN-DTGFLSFKDHL---PVTQIVITDT
              .     :.. .  :     :  . :. : .::... . .:   : . :   ..
gi|141 VRRGKSLKLTVD-DQQAMTGQM--AGDHTRLEFHNIETGIITERRYLSSVPSNFIGHLQS
          820        830         840       850       860       870

        810          820              830       840       850
FIRST_ DRSNSEAAW---RIGPL-RCYGDRRF------WNAVSFYTEASYLHFPTFHAEFSADISF
          :. :     . : .  :  . ::       : :.: :..::. . :..:  :  . :
gi|141 LTFNGMAYIDLCKNGDIDYCELNARFGFRNIIADPVTFKTKSSYVALATLQAYTSMHLFF
             880       890       900       910       920       930

        860       870        880       890       900       910
FIRST_ FFKTTALSGVFLENLGI-KDFIRLEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHY
        ::::.:.:..: : :  .::: .:. . . . ...:.:::   .  .: . :::::::
gi|141 QFKTTSLDGLILYNSGDGNDFIVVELVK-GYLHYVFDLGNGANLIKGSSNKPLNDNQWHN
             940       950        960       970       980       990

         920        930       940       950       960
FIRST_ VRAERNLKET-SLQVDNLPRSTRETSEEGHFRLQLNSQLFVGGTSSR-----------QK
       :   :. ..  ....:.  . : . .  :   :.:.:.:.::....            ..
gi|141 VMISRDTSNLHTVKIDT--KITTQITA-GARNLDLKSDLYIGGVAKETYKSLPKLVHAKE
             1000        1010       1020      1030      1040

           970       980       990       1000       1010      1020
FIRST_ GFLGCIRSLHLNGQKMDLEERAKVTSG-VRPGCPGHCSS-YGSICHNGGKCVEKHNGYLC
       :: ::. :. :::.  ::   :   .: .. :: :  ..   . : : : :... .:. :
gi|141 GFQGCLASVDLNGRLPDLISDALFCNGQIERGCEGPSTTCQEDSCSNQGVCLQQWDGFSC
      1050      1060      1070      1080      1090      1100

            1030       1040       1050      1060      1070
FIRST_ DCTNSPYEGPFCKKE-VSAVF-EAGTSVTYMFQEPYPVTKNISLSSSAIYTDSAPSKENI
       ::. . . ::.:.   .. .: ..: ..::     .:   :   :. :         . .
gi|141 DCSMTSFSGPLCNDPGTTYIFSKGGGQITY----KWP--PNDRPSTRA---------DRL
      1110      1120      1130          1140                 1150

    1080      1090      1100        1110      1120      1130
FIRST_ ALSFVTTQAPSLLLFINSSSQ--DFVVVLLCKNGSLQVRYHLNKEETHVFTIDADNFANR
```

```
       :..: :.:  ..:. ..:::   :.. . . ..:.. :..... ..  .   .:  . .
gi|141 AIGFSTVQKEAVLVRVDSSSGLGDYLELHI-HQGKIGVKFNVGTDDIAIEESNA-IINDG
           1160      1170      1180       1190      1200       1210

      1140      1150      1160      1170         1180      1190
FIRST_ RMHHLKINREGRELTIQMDQQLRLSYNFSPEVE-FRVIRSLTL--GKVTENLGLDSEVAK
       ..: ....: : . :.:.:.         : .: . . :.::.  ...:  .:     .:
gi|141 KYHVVRFTRSGGNATLQVDSW--------PVIERYPAGRQLTIFNSQATIIIG-----GK
             1220      1230              1240      1250

         1200      1210       1220             1230      1240
FIRST_ ANAMGFAGCMSSVQYNHIAPLK-AALRHATVAPV-------TVHGTLTESSCGFMVDSDV
        ... : : .:.. :: .  :. ::   :..: :        : ...:  : .  ..:..
gi|141 EQGQPFQGQLSGLYYNGLKVLNMAAENDANIAIVGNVRLVGEVPSSMTTESTATAMQSEM
      1260      1270      1280      1290      1300      1310

            1250      1260      1270      1280      1290      1300
FIRST_ NA-----VTTVHSSSDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMTRFL
       ..     .::. .:.   ::   .::...
gi|141 STSIMETTTTLATSTARRGKPPTKEPISQTTDDILVASAECPSDDEDIDPCEPSSGGLAN
      1320      1330      1340      1350      1360      1370


1345 residues in 1 query   sequences
1477 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:09:13 2002 done: Mon Dec 16 15:09:14 2002
 Scan time:  0.067 Display time:  2.133

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAAPEaigH: 1345 aa
 >FIRST_SEQUENCE
 vs  /tmp/fastaDAAQEaigH library
searching /tmp/fastaDAAQEaigH library

   1642 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.017
The best scores are:                                      opt
gi|21166380|ref|NP_620060.1| neurexin 2, isoform   (1642)  408

>>gi|21166380|ref|NP_620060.1| neurexin 2, isoform alpha  (1642 aa)
 initn: 339 init1: 146 opt: 408
Smith-Waterman score: 816;  24.250% identity in 1134 aa overlap (221-1279:265-1315)

              200       210       220       230       240       250
FIRST_ RFVRFVPLEWNPSGKIGMRVEVYGCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLKF
                                     :: : :   . : .... ...  : :.: :
gi|211 GFGGKFCSEEEHPMEGPAHLTLNSEGKEEFVATFKGNEFFCYDLSHNPIQSSTDEITLAF
          240       250       260       270       280       290

              260       270       280       290       300       310
FIRST_ KSMQGDGVLFHGEGQRGDHITLELQKGRLALHLNLGDSKARLSSSLPSATLGSLLDDQHW
       ...: .:...:  :. .:...: :..: . : .::: : :  .   :   ... ..:. :
gi|211 RTLQRNGLMLH-TGKSADYVNLSLKSGAVWLVINLG-SGAFEALVEP---VNGKFNDNAW
          300        310       320        330          340

               320                330       340       350
FIRST_ H-VLIERVGKQ--------VNFTVDK-HTQHFRTKGETDALDIDYELSFGGIP-VPGKPG
       : : . :  .:        :...::   :    :. .   :  :  . .:: : .   ::
gi|211 HDVRVTRNLRQHAGIGHAMVTISVDGILTTTGYTQEDYTMLGSDDFFYIGGSPNTADLPG
     350       360       370       380       390       400

     360       370       380            390       400        410
FIRST_ TFLKKNFHGCIENLYYNGVNIIL-----AKRRKHQIYTVGNVTFSCSE-PQIVPITFNSS
       . ...:: ::.... :.. .. :     ::.   ..   :...: : .   . :.::.:
gi|211 SPVSNNFMGCLKDVVYKNNDFKLELSRLAKEGDPKMKLQGDLSFRCEDVAALDPVTFESP
     410       420       430       440       450       460

           420       430       440       450                   460
FIRST_ GSYLLLPGTPQIDGLSVSFQFRTWNKDGLLLSTELSE---GSGT---------LLLSLEG
        ... ::        :.:..::: . .:::: ..  .   :.:.         . . :
gi|211 EAFVALPRWSAKRTGSISLDFRTTEPNGLLLFSQGRRAGGGAGSHSSAQRADYFAMELLD
     470       480       490       500       510       520

             470       480       490       500       510
FIRST_ GILRLVIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAP--PAPDSTWV
       : : :...  .  .    .. ..::: :  :...    . .........:   . ::  .
gi|211 GHLYLLLDMGSGGIKLRASSRKVNDGEWCHVDFQRDGRKGSISVNSRSTPFLATGDSEIL
     530       540       550       560       570       580

     520       530       540              550       560         570
FIRST_ QIYSGNSYYFGGCPDNLTDSQCLNP------IKA-FQGCMRLIFIDNQPKDL--ISVQQG
```

```
       .. :    :.:: :..   .  :  :      ..: . ::.: .:::.. .::  ..  ::
gi|211 DLES--ELYLGGLPEGGRVDLPLPPEVWTAALRAGYVGCVRDLFIDGRSRDLRGLAEAQG
     590         600       610       620       630       640

              580         590       600       610       620
FIRST_ SLGNFSDLHIDLCSIKD--RCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQ
       ..:        .:: .   .:    :..:: : ..:. : :.:  :.. : .:     :.
gi|211 AVGV-----APFCSRETLKQCASAPCRNGGVCREGWNRFICDCIGTGFLGRVC-----ER
       650            660       670       680       690

      630       640       650       660       670       680
FIRST_ SCEVYRHQGNTAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEK
          :  ..:.   .. :       . :  .. .  . ..    :.    .  . . .
gi|211 EATVLSYDGSM--YMKI-------MLPNAMHTEAEDVSLRFMSQRAYGLMMATTSRESAD
       700         710                720       730       740

      690       700       710            720              730
FIRST_ PYAMALDYGGSMEQLEAVIDGSE-----HCEQEVAYHC----RRSRLLN-TPDGTPFTWW
          . :: ::.:.    .  : :     :  ..  .:     ::.. :. . :..
gi|211 TLRLELD-GGQMKLTVNLGKGPETLFAGHKLNDNEWHTVRVVRRGKSLQLSVDNVTVEGQ
      750        760       770       780       790       800

      740       750       760       770       780       790
FIRST_ IGRSNERHPYWGGSPPGVQQCECGLDESCLDIQHFCNCDADKDEWTNDTGFLSFKDHLPV
       .. .. :  .        .. : :.    .  ..: .  . ..   . .: : :. .  .
gi|211 MAGAHMRLEF--------HNIETGI----MTERRFISV-VPSNFIGHLSG-LVFNGQPYM
       810               820           830        840        850

      800       810       820       830       840       850
FIRST_ TQIVITDTDRSNSEAAWRIGPLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADISFFF
        :    : : .  :   :.: ::        . :.: ...::: . :..:  :  . : :
gi|211 DQC--KDGDITYCELNARFG-LRAI----VADPVTFKSRSSYLALATLQAYASMHLFFQF
             860       870            880       890       900

      860       870        880       890       900       910
FIRST_ KTTALSGVFLENLGI-KDFIRLEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWHYVR
       :::: .:..: : :  .::: .:. . . : ...:.::::  .  .: . .:::::: :
gi|211 KTTAPDGLLLFNSGNGNDFIVIELVK-GYIHYVFDLGNGPSLMKGNSDKPVNDNQWHNVV
        910       920       930       940       950       960

       920        930       940       950                    960
FIRST_ AERNLKET-SLQVDNLPRSTRETSEEGHFRLQLNSQLFVGGTS-----------SRQKGF
       . :.  .. .:..:.  :.. . :. :   :.:...:..:: :           . . ::
gi|211 VSRDPGNVHTLKIDS--RTVTQHSN-GARNLDLKGELYIGGLSKNMFSNLPKLVASRDGF
        970       980          990      1000      1010      1020

         970       980       990             1000      1010      1020
FIRST_ LGCIRSLHLNGQKMDLEERAKVTSG-VRPGCPGH---CSSYGSICHNGGKCVEKHNGYLC
        ::. :. :::.  ::   :    : :. :: :    :.  .  : : : :... .:. :
gi|211 QGCLASVDLNGRLPDLIADALHRIGQVERGCDGPSTTCTEES--CANQGVCLQQWDGFTC
          1030      1040      1050      1060        1070      1080

            1030      1040      1050      1060      1070
FIRST_ DCTNSPYEGPFCKKEVSAVFEAGTSVTYMFQEPYPVTKNISLSSSAIYTDSAPSK--ENI
       ::: . : :: :.       . ::  ::.:       :. .: . .   .. ::   . .
gi|211 DCTMTSYGGPVCN-------DPGT--TYIFG------KGGALITYTWPPNDRPSTRMDRL
            1090               1100            1110      1120

    1080      1090        1100      1110      1120      1130
FIRST_ ALSFVTTQAPSLLLFINSSS--QDFVVVLLCKNGSLQVRYHLNKEETHVFTIDADN--FA
```

```
       :..: : :  ..:. ..:.:   :.. . .  .:.. : .... ..   .:::  :   .
gi|211 AVGFSTHQRSAVLVRVDSASGLGDYLQLHI-DQGTVGVIFNVGTDD---ITIDEPNAIVS
        1130      1140      1150       1160      1170         1180

        1140      1150      1160      1170      1180      1190
FIRST_ NRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLTLGKVTENLGLDSEVAKA
       . ..: ....: : . :.:.:.   ..  .    .. .. : .  :.    : :.
gi|211 DGKYHVVRFTRSGGNATLQVDSW-PVNERYPAGRQLTIFNSQAAIKIG---GRDQ-----
            1190      1200       1210      1220         1230

        1200      1210      1220      1230      1240      1250
FIRST_ NAMGFAGCMSSVQYNHIAPLKAALRHATVAPVTVHGTLTESSCGFMVDSDVNAVTTVHSS
        .  : : .:.. :: .  :  : .    . .  :  :.  . . ......:.: .. .
gi|211 -GRPFQGQVSGLYYNGLKVLALAAESDPNVRTEGHLRLVGEGPSVLLSAETTATTLLADM
            1240      1250      1260      1270      1280      1290

        1260      1270      1280      1290      1300      1310
FIRST_ SDPFGKTDEREPLTNAVRSDSAVIGGVIAVVIFIIFCIIGIMTRFLYQHKQSHRTSQMKE
        . . .:      :.. :. : ..
gi|211 ATTIMETTTTMATTTTRRGRSPTLRDSTTQNTDDLLVASAECPSDDEDLEECEPSTGGEL
            1300      1310      1320      1330      1340      1350


1345 residues in 1 query   sequences
1642 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:05:13 2002 done: Mon Dec 16 15:05:15 2002
 Scan time:  0.017 Display time:  2.166

Function used was FASTA
```

```
 FASTA searches a protein or DNA sequence data bank
 version 3.3t05 March 30, 2000
Please cite:
 W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

/tmp/fastaCAA4caW3G: 1345 aa
 >FIRST_SEQUENCE
 vs  /tmp/fastaDAA5caW3G library
searching /tmp/fastaDAA5caW3G library

   1392 residues in     1 sequences

FASTA (3.34 January 2000) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 40, opt: 28, gap-pen: -12/ -2, width:  16
 Scan time:  0.033
The best scores are:                                      opt
gi|23498650|emb|CAC87720.2| neurexin 3-alpha [Hom  (1392)  369

>>gi|23498650|emb|CAC87720.2| neurexin 3-alpha [Homo sap  (1392 aa)
 initn: 324 init1: 139 opt: 369
Smith-Waterman score: 774;  25.791% identity in 1012 aa overlap (220-1156:255-1184)

     190       200       210       220       230       240
FIRST_ ARFVRFVPLEWNPSGKIGMRVEVYGCSYKSDVADFDGRSSLLYRFNQKLMSTLKDVISLK
                                     .:: : :   : : ..:. ... .: :.:.
gi|234 STTGYGGKLCSEGLSHLMMSEQGRSKAREENVATFRGSEYLCYDLSQNPIQSSSDEITLS
          230       240       250       260       270       280

     250       260       270       280       290       300
FIRST_ FKSMQGDGVLFHGEGQRGDHITLELQKGRLALHLNLGDSKARLSSSLPSATLGSLLDDQH
       ::. : .:...:  :. .:...: :. : ..: .::: : :  .   :   ... ..:.
gi|234 FKTWQRNGLILH-TGKSADYVNLALKDGAVSLVINLG-SGAFEAIVEP---VNGKFNDNA
          290        300       310       320        330

     310        320        330       340       350        360
FIRST_ WH-VLIERVGKQVNFTVDK-HTQHFRTKGETDALDIDYELSFGGIPVPGK-PGTFLKKNF
       :: : . :  .::..::    :    :. .   :  :  .  :: :  .  ::. ...::
gi|234 WHDVKVTRNLRQVTISVDGILTTTGYTQEDYTMLGSDDFFYVGGSPSTADLPGSPVSNNF
     340       350       360       370       380       390

        370       380              390       400        410
FIRST_ HGCIENLYYNGVNIILAKRR-------KHQIYTVGNVTFSCSE-PQIVPITFNSSGSYLL
        ::.... :.. .: :   :       : .::  :.:.:.: .   . ::.:..  .:.
gi|234 MGCLKEVVYKNNDIRLELSRLARIADTKMKIY--GEVVFKCENVATLDPINFETPEAYIS
     400       410       420       430         440       450

      420       430       440                 450       460
FIRST_ LPGTPQIDGLSVSFQFRTWNKDGLLLSTE----------LSEGSGTLLLSLE--GGILRL
       ::.       :.:::.:: . .::.: :.           .....  ....:   : : :
gi|234 LPKWNTKRMGSISFDFRTTEPNGLILFTHGKPQERKDARSQKNTKVDFFAVELLDGNLYL
       460       470       480       490       500       510

        470       480       490       500       510        520
FIRST_ VIQKMTERVAEILTGSNLNDGLWHSVSINARRNRITLTLDDEAAP-PAPDSTWVQIYSGN
       ...  .  .    : .. ::: :. :.:.      :...... .:  :   . .    :.
gi|234 LLDMGSGTIKVKATQKKANDGEWYHVDIQRDGRSGTISVNSRRTPFTASGESEILDLEGD
       520       530       540       550       560       570

         530           540         550       560         570
FIRST_ SYYFGGCPDN----LTDSQCLNPI--KAFQGCMRLIFIDNQPKDL--ISVQQGSLGNFSD
```

```
        .:.:: :.:    .  ..  . .   .. ::.: .::::.. :..  .. .:.. :  :.
gi|234 -MYLGGLPENRAGLILPTELWTAMLNYGYVGCIRDLFIDGRSKNIRQLAEMQNAAGVKSS
        580       590       600       610       620       630

       580         590       600       610       620       630
FIRST_ LHIDLCS--IKDRCLPNYCEHGGSCSQSWTTFYCNCSDTSYTGATCHNSIYEQSCEVYRH
            ::     .:    :.... :...:. : :.:. :.: : ::.   : :   :
gi|234 -----CSRMSAKQCDSYPCKNNAVCKDGWNRFICDCTGTGYWGRTCER---EASILSY--
             640       650       660       670          680

         640       650       660       670       680       690
FIRST_ QGNTAGFFYIDSDGSGPLGPLQVYCNITEDKIWTSVQHNNTELTRVRGANPEKPYAM--A
                   :::       .: .:    :   :.:...: .  : . .. :..  :
gi|234 ------------DGS-------MYMKI----IMPMVMHTEAEDVSFRFMS-QRAYGLLVA
                           690           700       710        720

           700       710       720       730       740
FIRST_ LDYGGSMEQLEAVIDGSEHCEQEVAYHCRRSRLLNTPDGTPFTWWIGR---SNERHPY--
            : . :.  .::.           : . ..:   : : : . :.   .:: :
gi|234 TTSRDSADTLRLELDGG-----------RVKLMVNLGKG-PETLYAGQKLNDNEWHTVRV
            730                  740       750        760       770

        750            760       770       780         790
FIRST_ --WGGS-----PPGVQQCECGLDESCLDIQHFCNCDADKDEWTN--DTGFLSFKDHLPVT
          : :        : .     :.. :..... .    . .. .   ..:..  . :  .
gi|234 VRRGKSLKLTVDDDVAEGTMVGDHTRLEFHNIETGIMTEKRYISVVPSSFIGHLQSLMFN
              780       790       800       810       820       830

     800           810       820       830       840       850
FIRST_ QIVITDT----DRSNSEAAWRIGPLRCYGDRRFWNAVSFYTEASYLHFPTFHAEFSADIS
        ..  :     : .  :   :.: ::      . . :.: :..::: . :..:  :  .
gi|234 GLLYIDLCKNGDIDYCELKARFG-LRNI----IADPVTFKTKSSYLSLATLQAYTSMHLF
              840       850            860       870       880

         860       870        880       890       900       910
FIRST_ FFFKTTALSGVFLENLGI-KDFIRLEISSPSEITFAIDVGNGPVELVVQSPSLLNDNQWH
       : ::::. .: .: : :  .::: .:. . . : ...:.::::  .  .:   :::::::
gi|234 FQFKTTSPDGFILFNSGDGNDFIAVELVK-GYIHYVFDLGNGPNVIKGNSDRPLNDNQWH
         890       900       910        920       930       940

          920        930       940          950                  960
FIRST_ YVRAERNLKET-SLQVDNLPRSTRETSE--EGHFRLQLNSQLFVGGTS-----------S
        :   :....: ::.::     :. ...  .:   :.:...:...: .           .
gi|234 NVVITRDNSNTHSLKVD-----TKVVTQVINGAKNLDLKGDLYMAGLAQGMYSNLPKLVA
          950       960            970       980       990

              970       980       990       1000       1010
FIRST_ RQKGFLGCIRSLHLNGQKMDLEERAKVTSG-VRPGCPGHCSS-YGSICHNGGKCVEKHNG
        . :: ::. :. :::.  :: . :   :: .. :: :  ..   . : : : :... .:
gi|234 SRDGFQGCLASVDLNGRLPDLINDALHRSGQIERGCEGPSTTCQEDSCANQGVCMQQWEG
    1000      1010      1020      1030      1040      1050

     1020      1030       1040       1050      1060      1070
FIRST_ YLCDCTNSPYEGPFCKKE-VSAVF-EAGTSVTYMFQEPYPVTKNISLSSSAIYTDSAPSK
       . :::. . : : :.   .. .: ..:  . :     .:..   :  :
gi|234 FTCDCSMTSYSGNQCNDPGATYIFGKSGGLILYT----WPANDRPSTRS-----------
    1060      1070      1080      1090          1100

       1080      1090      1100        1110      1120      1130
FIRST_ ENIALSFVTTQAPSLLLFINSSSQ--DFVVVLLCKNGSLQVRYHLNKEETHVFTIDADNF
```

```
       . .:..: ::   ..:. :.:.    ::. . . ..:.. : ....  .  .   .
gi|234 DRLAVGFSTTVKDGILVRIDSAPGLGDFLQLHI-EQGKIGVVFNIGTVDISIKE-ERTPV
         1110      1120      1130      1140      1150      1160

         1140      1150      1160      1170      1180      1190
FIRST_ ANRRMHHLKINREGRELTIQMDQQLRLSYNFSPEVEFRVIRSLTLGKVTENLGLDSEVAK
        . ..: ....:.: . :.:.:.
gi|234 NDGKYHVVRFTRNGGNATLQVDNWPVNEHYPTGNTDNERFQMVKQKIPFKYNRPVEEWLQ
           1170      1180      1190      1200      1210      1220


1345 residues in 1 query   sequences
1392 residues in 1 library sequences
 Scomplib [version 3.3t05 March 30, 2000]
 start: Mon Dec 16 15:07:25 2002 done: Mon Dec 16 15:07:26 2002
 Scan time:  0.033 Display time:  1.817

Function used was FASTA
```

characterize the protein. A starting material that can only be used to produce a final product does not have a substantial asserted utility in those instances where the final product is not supported by a specific and substantial utility. In this case none of the proteins that are to be produced as final products resulting from processes involving the claimed cDNA have asserted or identified specific and substantial utilities. The research contemplated by Applicants to characterize potential protein products, especially their biological activities, does not constitute a specific and substantial utility. Identifying and studying the properties of the protein itself or the mechanisms in which the protein is involved does not define a "real world" context of use. Note, because the claimed invention is not supported by a specific and substantial asserted utility for the reasons set forth above, credibility has not been assessed. Neither the specification as filed nor any art of record discloses or suggests any property or activity for the cDNA compounds such that another non-asserted utility would be well established for the compounds.

Claim 1 is also rejected under 35 U.S.C. § 112, first paragraph. Specifically, since the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art would not know how to use the claimed invention.

Example 10: DNA Fragment encoding a Full Open Reading Frame (ORF)

Specification: The specification discloses that a cDNA library was prepared from human kidney epithelial cells and 5000 members of this library were sequenced and open reading frames were identified. The specification discloses a Table that indicates that one member of the library having SEQ ID NO: 2 has a high level of homology to a DNA ligase. The specification teaches that this complete ORF (SEQ ID NO: 2) encodes SEQ ID NO: 3. An alignment of SEQ ID NO: 3 with known amino acid sequences of DNA ligases indicates that there is a high level of sequence conservation between the various known ligases. The overall level of sequence similarity between SEQ ID NO: 3 and the consensus sequence of the known DNA ligases that are presented in the specification reveals a similarity score of 95%. A search of the prior art confirms that SEQ ID NO: 2 has high homology to DNA Ligase encoding nucleic acids and that the next highest level of homology is to alpha-actin. However, the latter homology is only 50%. Based on the sequence homologies, the specification asserts that SEQ ID NO: 2 encodes a DNA ligase.

Claim 1: An isolated and purified nucleic acid comprising SEQ ID NO: 2.

Analysis: The following analysis includes the questions that need to be asked according to the guidelines and the answers to those questions based on the above facts:

1) Based on the record, is there a "well established utility" for the claimed invention? Based upon applicant's disclosure and the results of the PTO search, there is no reason to doubt the assertion that SEQ ID NO: 2 encodes a DNA ligase. Further, DNA ligases have a well-established use in the molecular biology art based on this class of protein's ability to ligate DNA. Consequently the answer to the question is yes.

Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed. In order to determine whether the claimed invention has a well-established utility the examiner must determine that the invention has a specific, substantial and credible utility that would have been readily apparent to one of skill in the art. In this case SEQ ID NO: 2 was shown to encode a DNA ligase that the artisan would have recognized as having a specific, substantial and credible utility based on its enzymatic activity.

Thus, the conclusion reached from this analysis is that a 35 U.S.C. § 101 rejection and a 35 U.S.C. § 112, first paragraph, utility rejection should not be made.

Example 11: Animals with Uncharacterized Human Genes

Specification: Kidney cells from a patient with Polycystic Kidney (PCK) Disease have been used to make a cDNA library. From this library 8000 nucleotide "fragments" have been sequenced but not yet used to express proteins in a transformed host cell nor have they been characterized in any other way. The 50 longest fragments, SEQ ID NO: 1-50, respectively, have been used to make transgenic mice. None of the 50 lines of mice have developed Polycystic Kidney Disease to date. The asserted utility is the use of the mice to research human genes from diseased human kidneys. The disease is inheritable, but chromosomal loci have not yet been identified. Neither the absence or presence of a specific protein has been identified with the disease condition.

W

US005445934A

United States Patent [19]

Fodor et al.

[11] **Patent Number: 5,445,934**

[45] **Date of Patent: Aug. 29, 1995**

[54] **ARRAY OF OLIGONUCLEOTIDES ON A SOLID SUBSTRATE**

[75] Inventors: **Stephen P. A. Fodor**, Palo Alto, Calif.; **Michael C. Pirrung**, Durham, N.C.; **J. Leighton Read**, Palo Alto; **Lubert Stryer**, Stanford, both of Calif.

[73] Assignee: **Affymax Technologies N.V.**, Curacao, Netherlands Antilles

[21] Appl. No.: **954,646**

[22] Filed: **Sep. 30, 1992**

Related U.S. Application Data

[60] Division of Ser. No. 850,356, Mar. 12, 1992, which is a division of Ser. No. 492,462, Mar. 7, 1990, Pat. No. 5,143,854, which is a continuation-in-part of Ser. No. 362,901, Jun. 7, 1989, abandoned.

[51] **Int. Cl.**^{6} **C12Q 1/68**; G01N 33/543

[52] **U.S. Cl.** **435/6**; 435/7.92; 435/969; 435/973; 436/518; 436/527; 436/807; 436/809; 536/25.3; 428/426; 428/532

[58] **Field of Search** 435/7.92, 7.94, 7.95, 435/969, 973, 6; 536/25.3–25.34; 935/88; 436/518, 527, 807, 809; 530/334; 427/2; 428/436, 532

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,562,157	12/1985	Lowe	435/291
4,689,405	8/1987	Frank et al.	536/27
4,728,591	3/1988	Clark et al.	430/5
4,886,741	12/1989	Schgwartz	435/5
4,888,278	12/1989	Singer et al.	435/6

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

0328256	8/1989	European Pat. Off.	B01J 20/32
WO89/10977	11/1989	WIPO	C12Q 1/68
WO90/03382	5/1990	WIPO	C07H 21/00

OTHER PUBLICATIONS

BioRad Chromatography, Electrophoresis, Immunochemistry, Molecular Biology, HPLC Catalogue M 1987 p. 182.

Hames & Heggins (ed). Nucleic Acid Hybridization: A Practical Approach (1985) IRL Press Oxford England.

Science, 253, p. 1489, Sep. 27, 1991 "Will DNA Chip Speed Genome Initiative?"

Khrapko et al., FEB, 256, pp. 118–122, Oct. 1989 "An oligonucleotide hybridization approach to DNA sequencing".

Mirzabekov, Tibtech, 12, pp. 27–32, Jan. 1994 "DNA sequencing by hybridization—a megasequencing method and a diagnostic tool?"

Southern et al., Genomics, 13, pp. 1008–1017, 1992 "Analyzing and Comparing Nucleic Acid Sequences by Hybridization to Arrays of Oligonucleotides: Evaluation Using Experimental Models".

Haridasan et al., "Peptide Synthesis Using Photolyti-

(List continued on next page.)

Primary Examiner—Esther M. Kepplinger
Assistant Examiner—Lora M. Green
Attorney, Agent, or Firm—Townsend and Townsend Khourie and Crew

[57] **ABSTRACT**

A method and apparatus for preparation of a substrate containing a plurality of sequences. Photoremovable groups are attached to a surface of a substrate. Selected regions of the substrate are exposed to light so as to activate the selected areas. A monomer, also containing a photoremovable group, is provided to the substrate to bind at the selected areas. The process is repeated using a variety of monomers such as amino acids until sequences of a desired length are obtained. Detection methods and apparatus are also disclosed.

10 Claims, 20 Drawing Sheets

U.S. PATENT DOCUMENTS

4,923,901	5/1990	Koester et al.	536/25.3 X
4,973,493	11/1990	Guire	427/2
5,079,600	1/1992	Schnur et al.	357/4
5,143,854	9/1992	Pirrung et al.	436/518
5,202,231	4/1993	Drmanac et al.	435/6

OTHER PUBLICATIONS

cally Cleavable 2-Nitrobenzyloxycarbonyl Protecting Group," *Proc. Indian Natl. Sci. Acad., Part A* (1987) 53:717–728.

Sze/McGillis, *VLSI Technology,* Chapter 7, pp. 267–301, McGraw-Hill, 1983.

Geysen et al., "Strategies for epitope analysis using peptide synthesis," *J. Immunol. Meth.* (1987) 102:259–274.

Furka et al., "More Peptides by Less Labor," Abstract No. 288 from *Xth International Symposium on Medicinal Chemistry,* Budapest, Hungary, Aug. 15–19, 1988.

Ohtsuka et al., "Studies on transfer ribonucleic acids and related compounds. IX(1) Ribooligonucleotide synthesis . . . 2'-hydroxyl group," *Nucleic Acids Research* (1974) 1:1351–1357.

Kleinfeld et al., "Controlled Outgrowth of Dissociated Neurons on Patterned Substrates," *J. of Neuroscience,* 8(11):4098–4120, Nov. 1988.

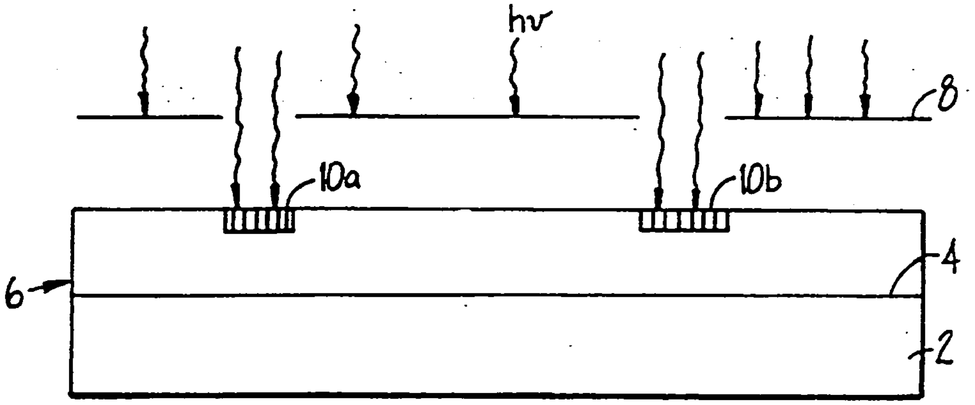


FIG._1.

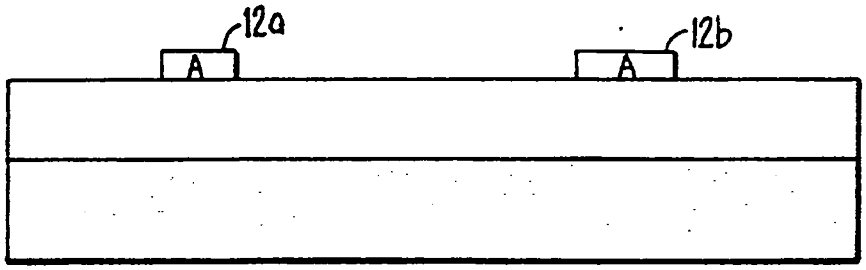


FIG._2.

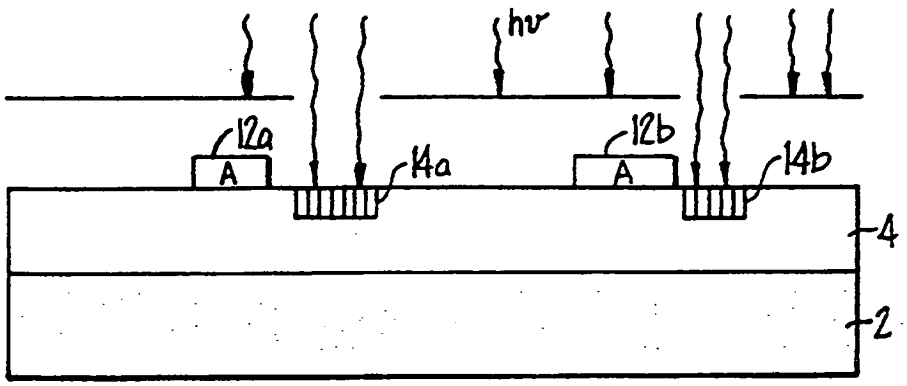


FIG._3.

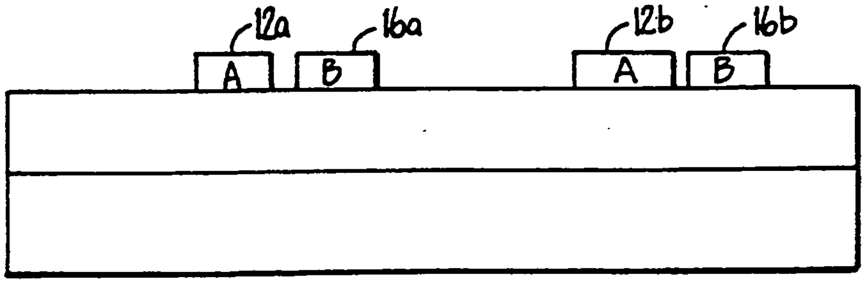


FIG._4.

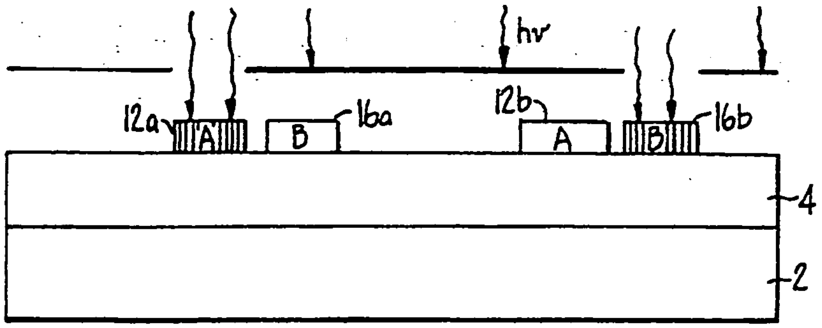


FIG._5.

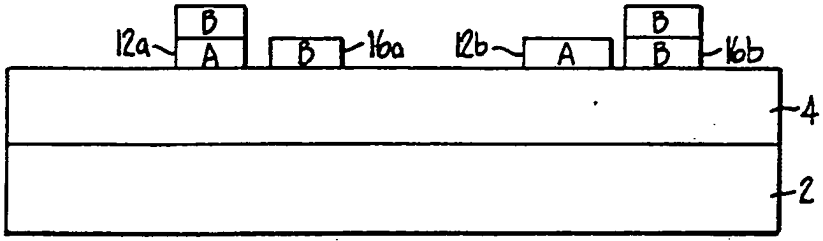


FIG._6.

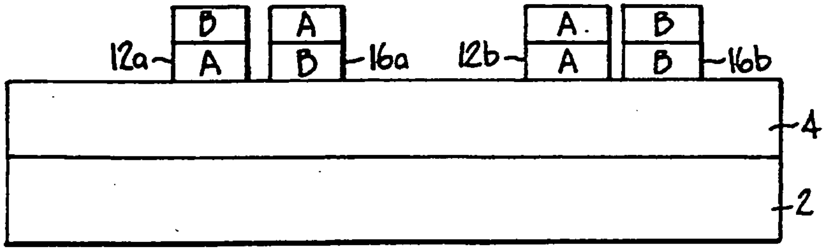


FIG._7.

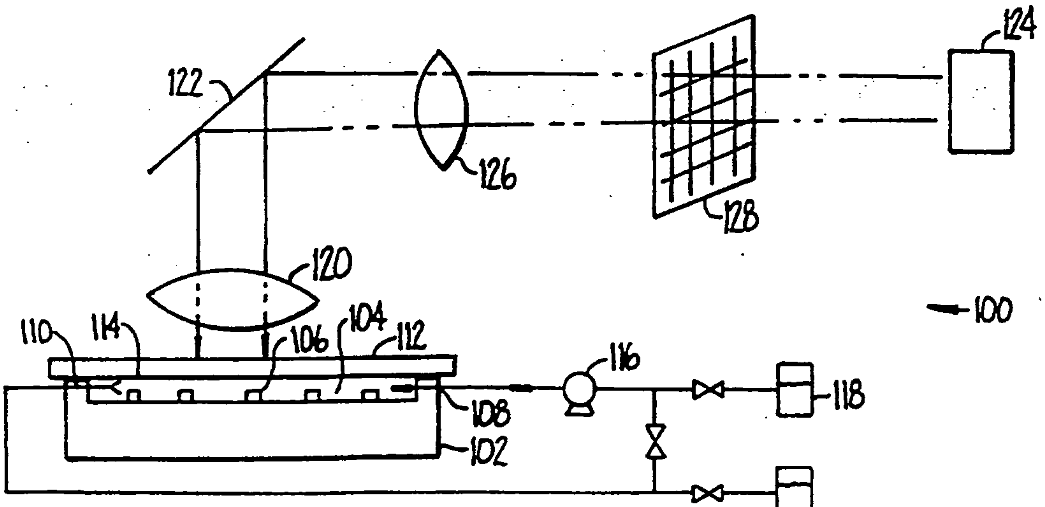


FIG._8a.

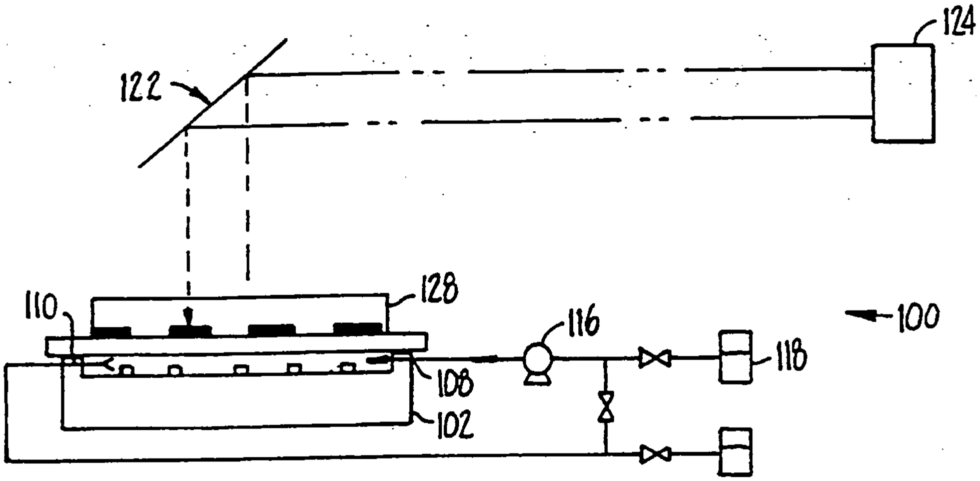


FIG._8b.

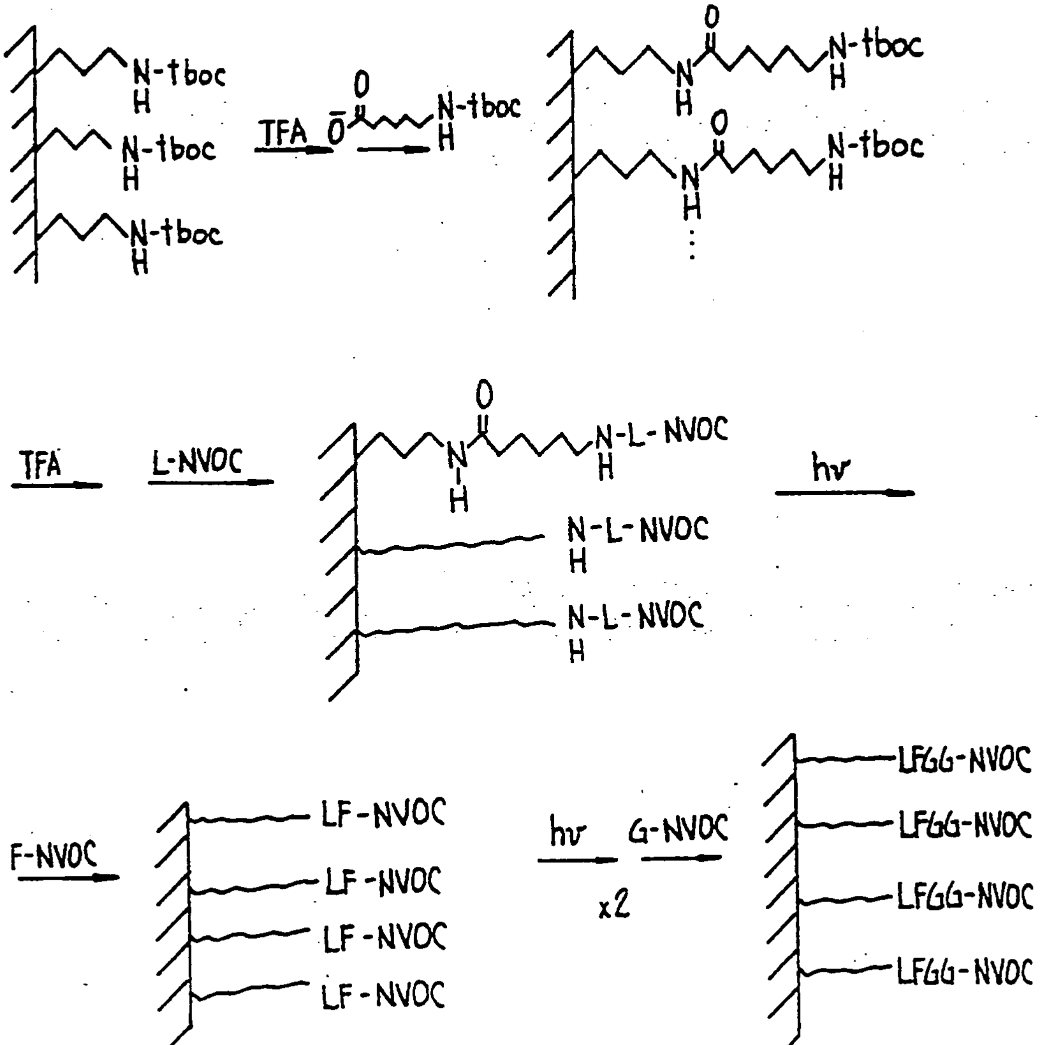


FIG._14A.

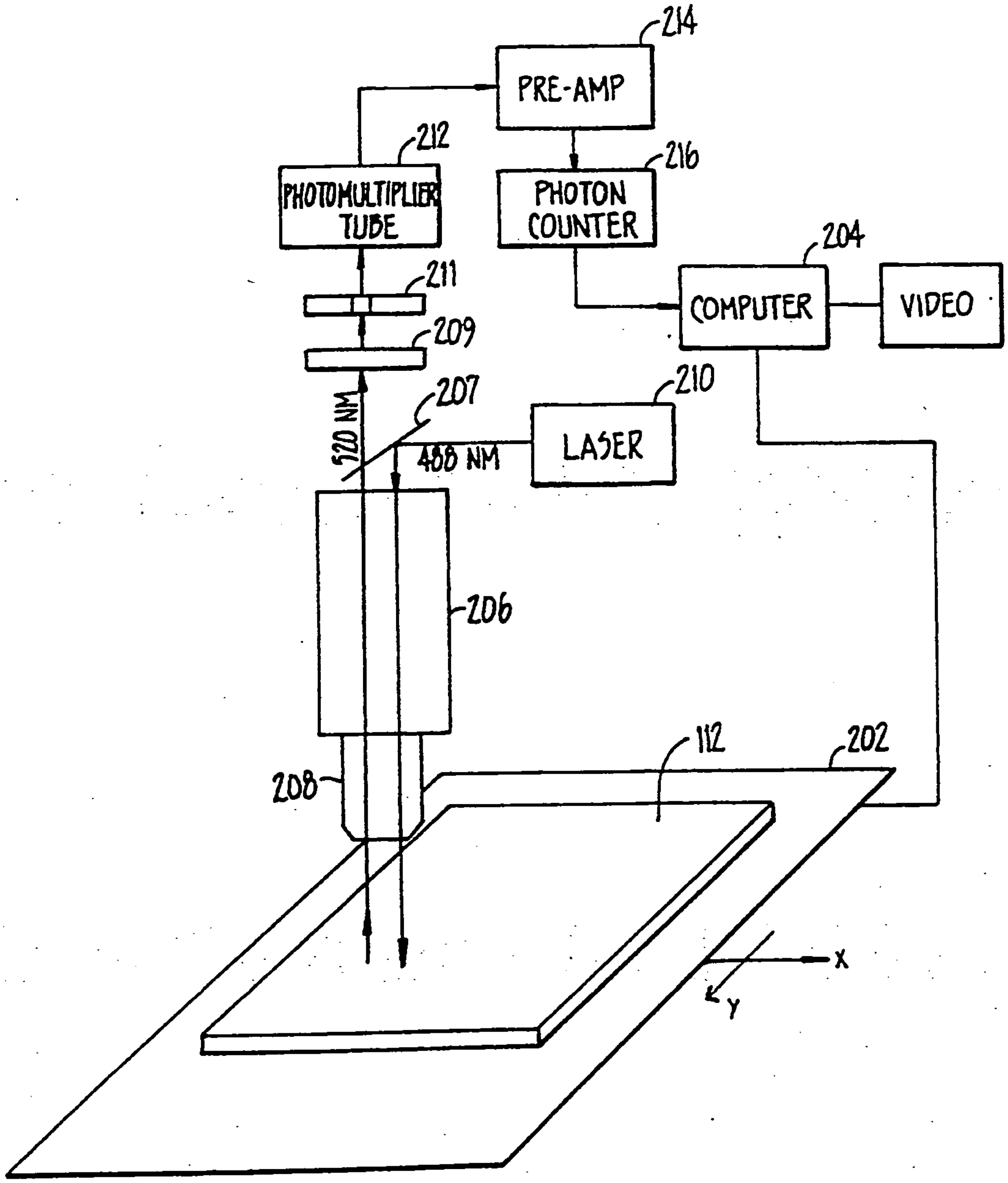

214
PRE-AMP
212
PHOTOMULTIPLIER TUBE
216
PHOTON COUNTER
204
COMPUTER
VIDEO
211
209
207
210
LASER
520 NM
488 NM
206
112
202
208
X
Y

FIG._9.

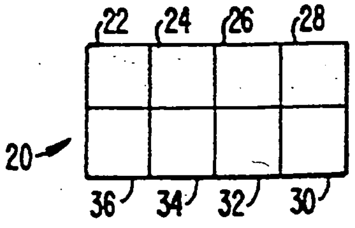


FIG._10A.

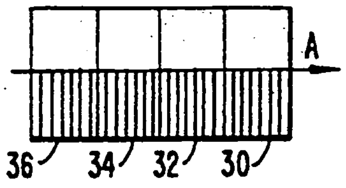
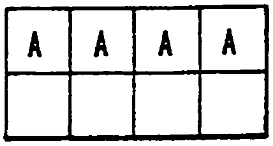


FIG._10B.



A	A	A	A

FIG._10C.

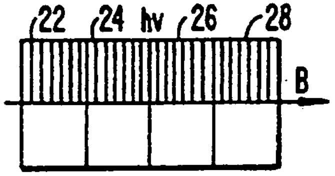
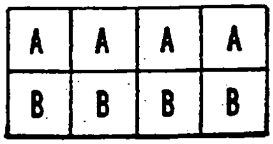


FIG._10D.



A	A	A	A
B	B	B	B

FIG._10E.

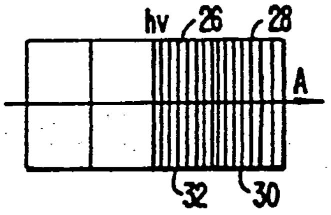
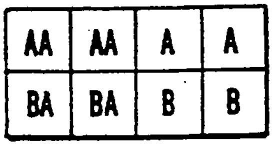


FIG._10F.



AA	AA	A	A
BA	BA	B	B

FIG._10G.

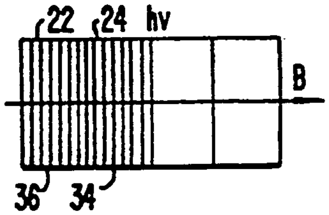
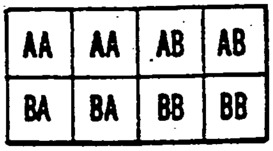


FIG._10H.



AA	AA	AB	AB
BA	BA	BB	BB

FIG._10I.

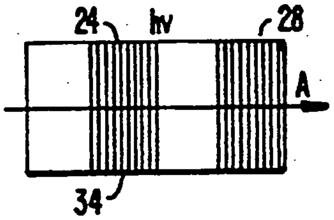
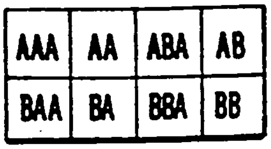


FIG._10J.



AAA	AA	ABA	AB
BAA	BA	BBA	BB

FIG._10K.

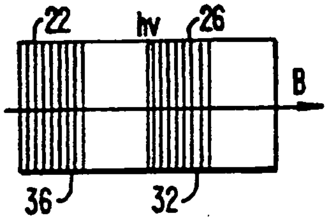
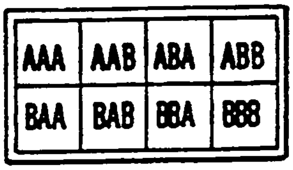
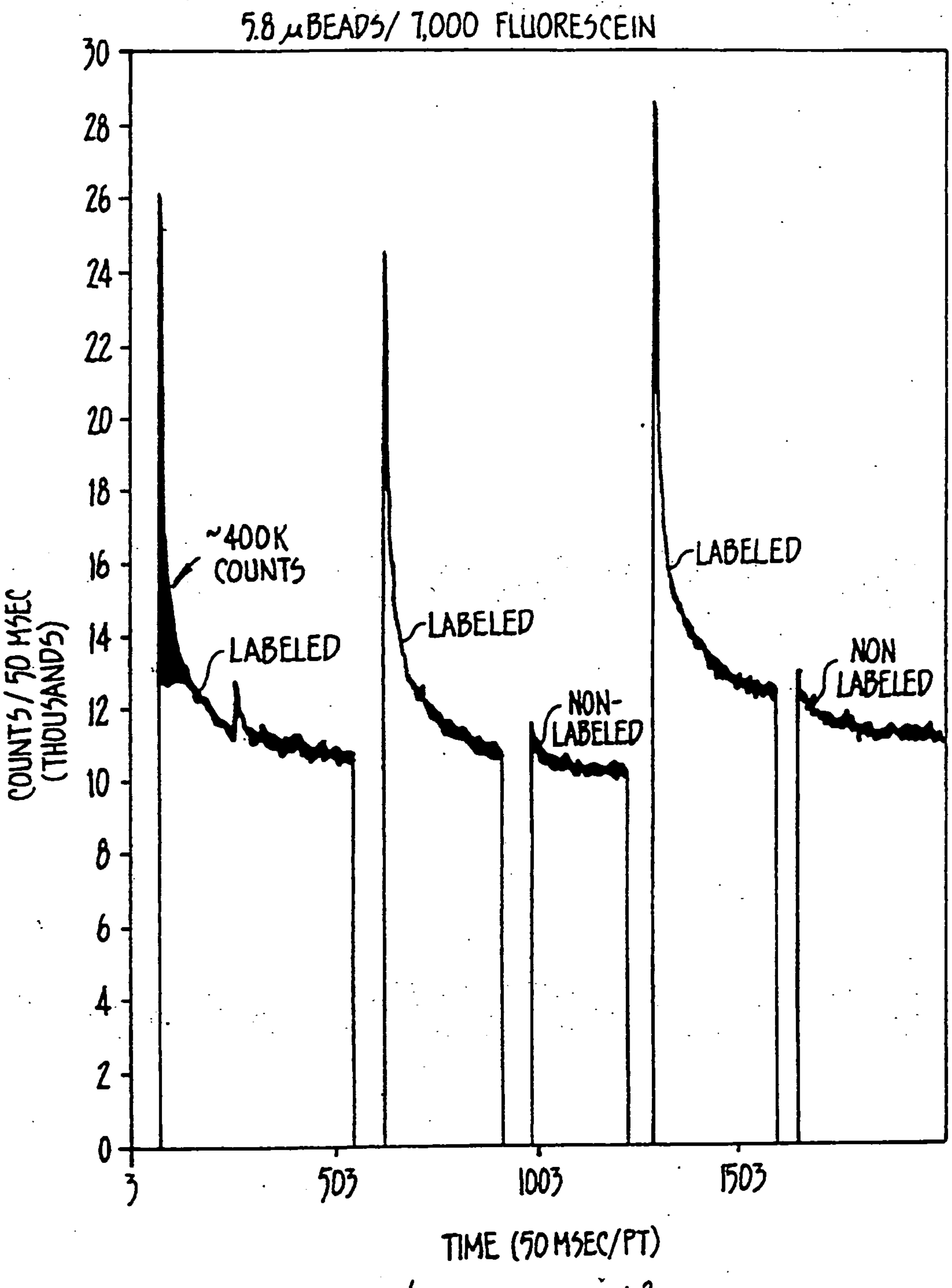


FIG._10L.



AAA	AAB	ABA	ABB
BAA	BAB	BBA	BBB

FIG._10M.



~2 x 10^{-6} CHROMOPHORE/A^2

FIG._11A.

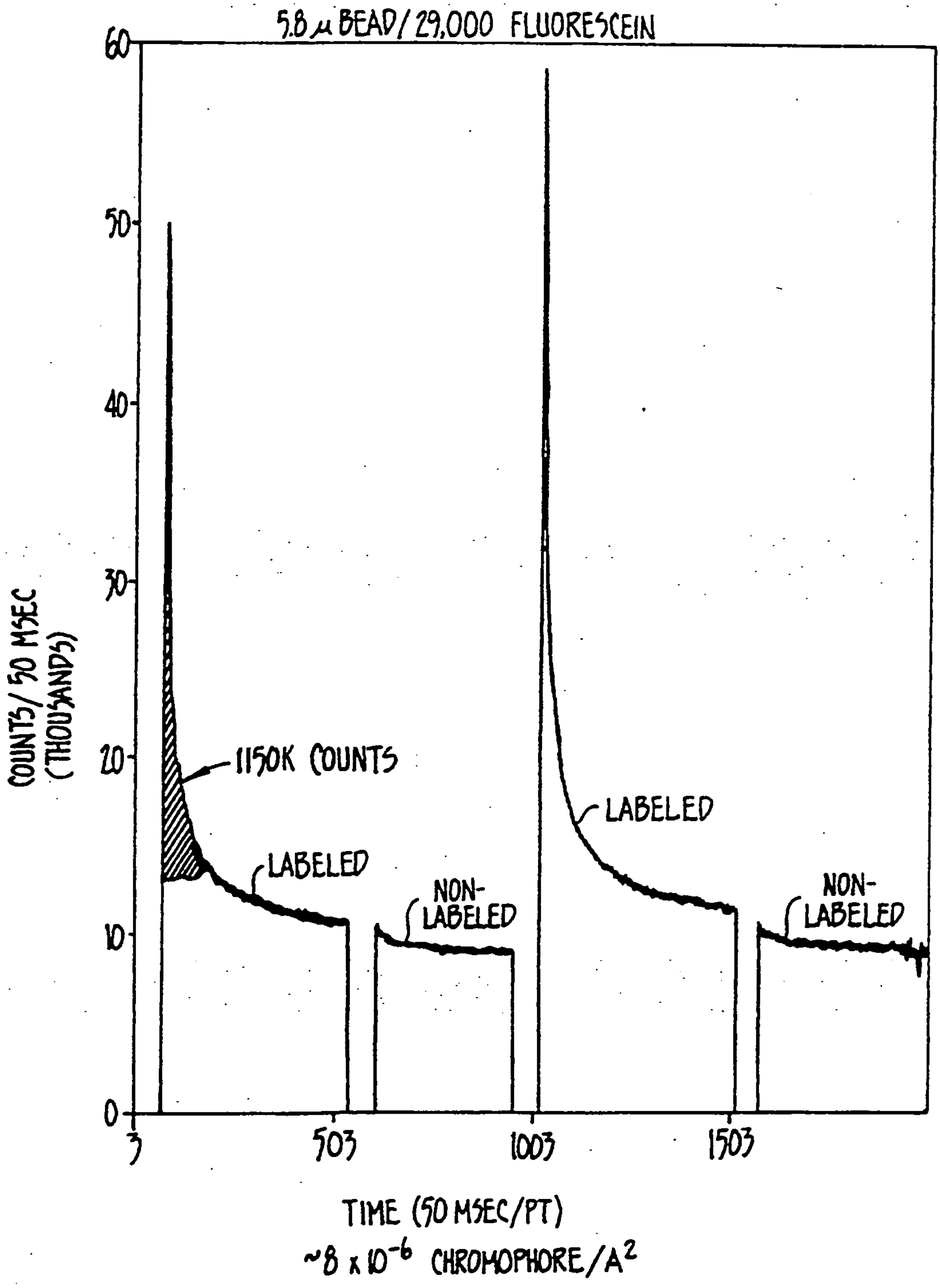

5.8 μ BEAD/29,000 FLUORESCEIN
60
50
40
30
20
10
0
COUNTS/50 MSEC (THOUSANDS)
1150K COUNTS
LABELED
NON-LABELED
LABELED
NON-LABELED
3
503
1003
1503
TIME (50 MSEC/PT)
~8 x 10^-6 CHROMOPHORE/A^2

FIG._11B.

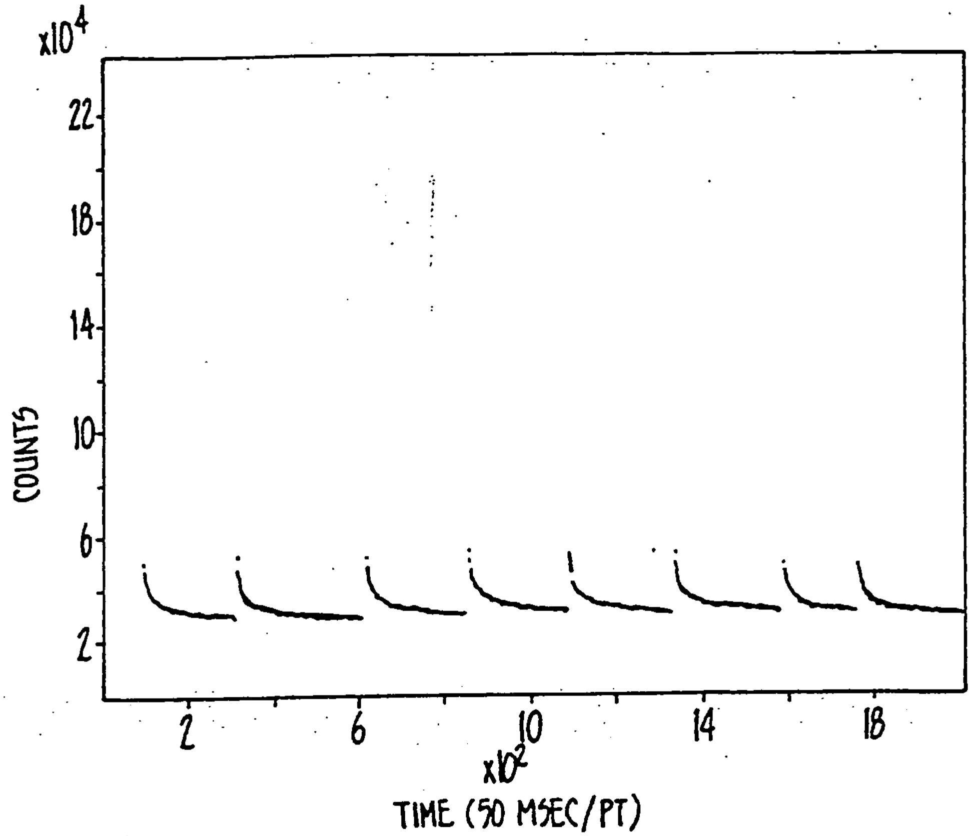


FIG_12A.

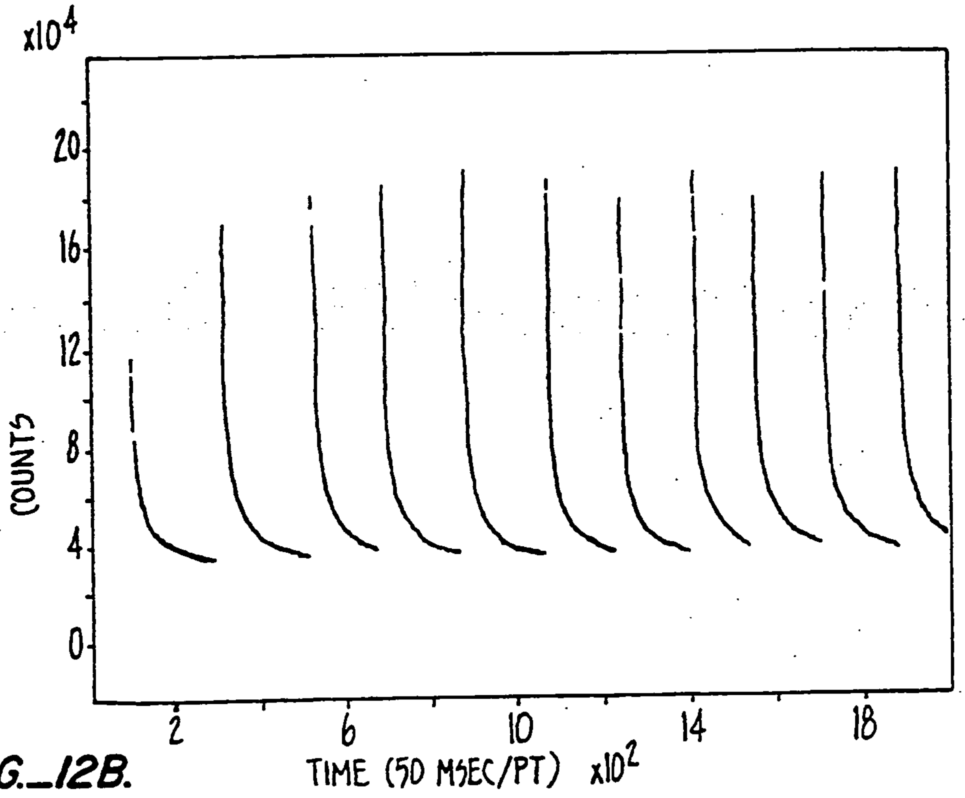
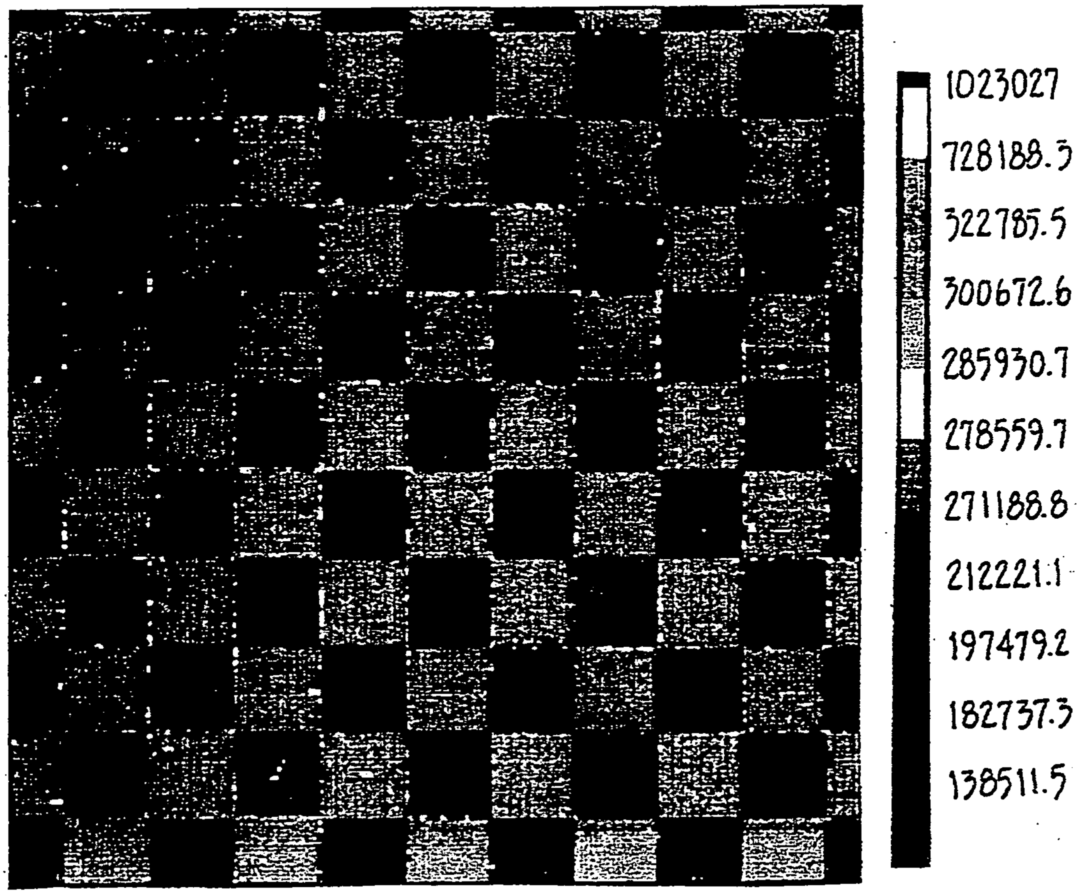
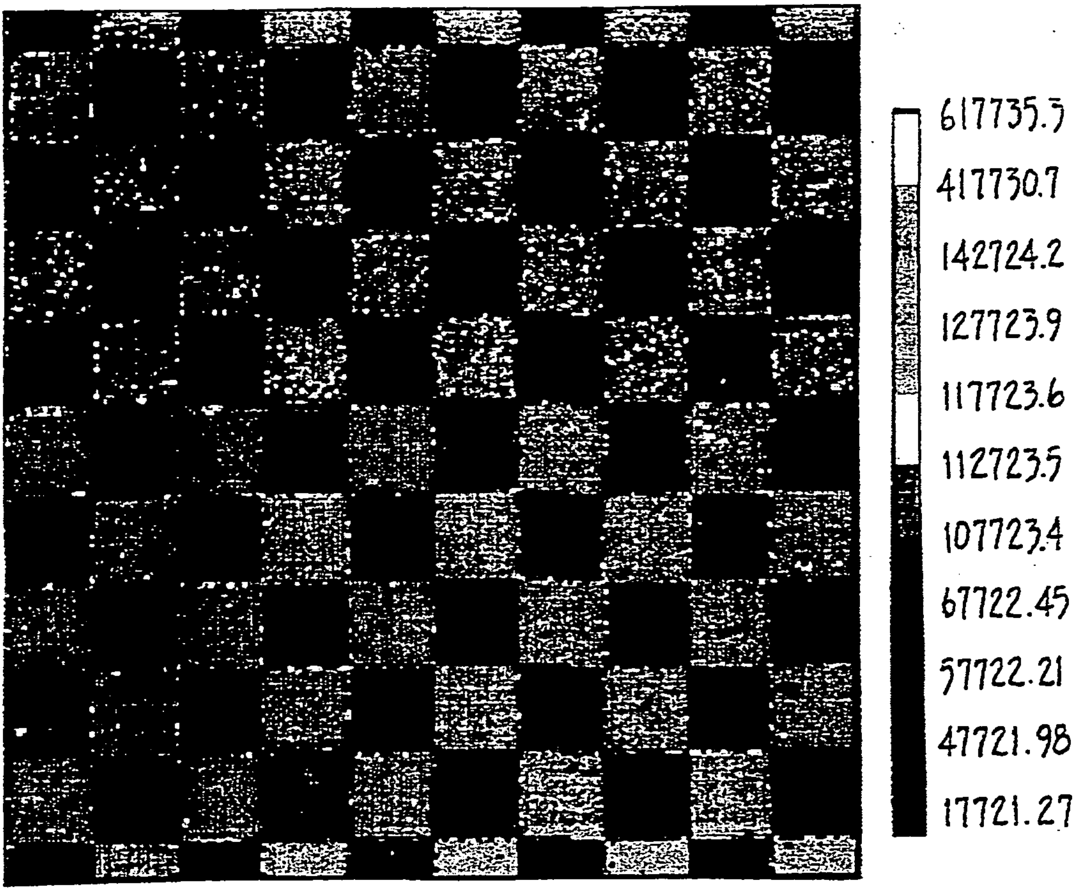


FIG_12B.



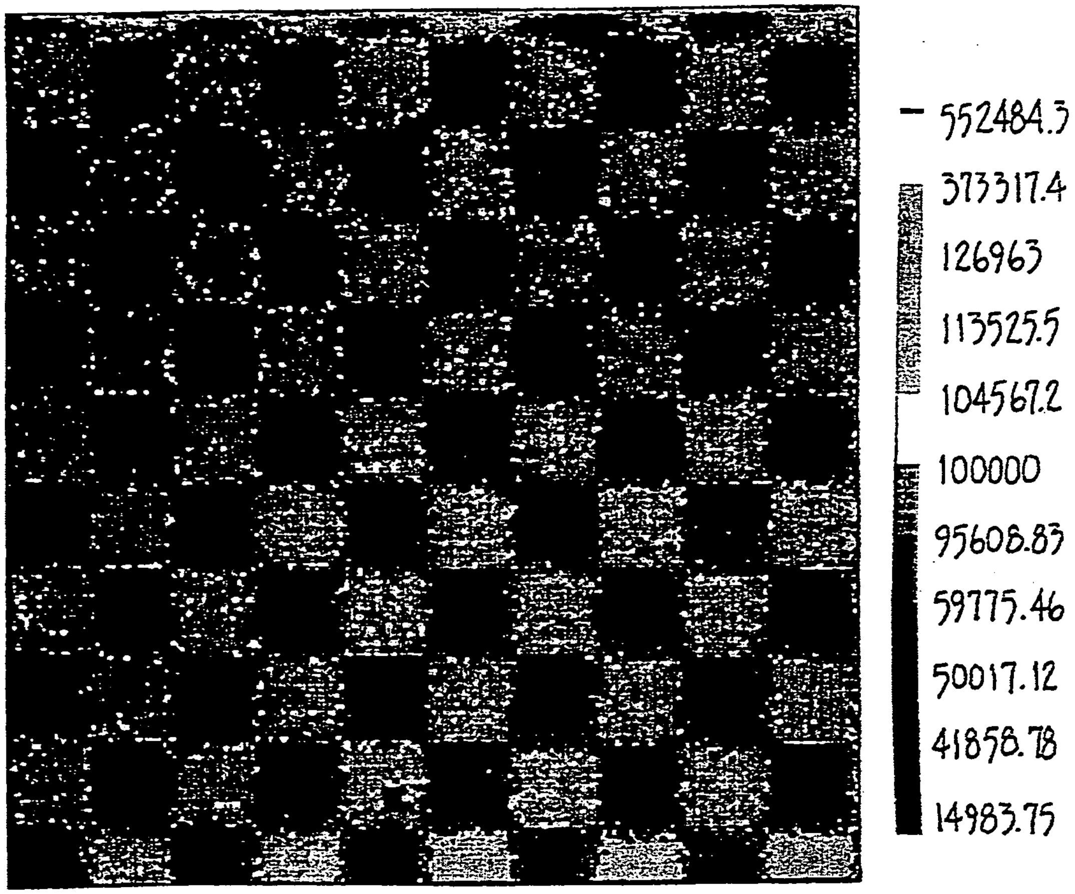
MEAN: 285930.7
VAR: 2.173242E+10
σ: 147419.2

FIG._13A.



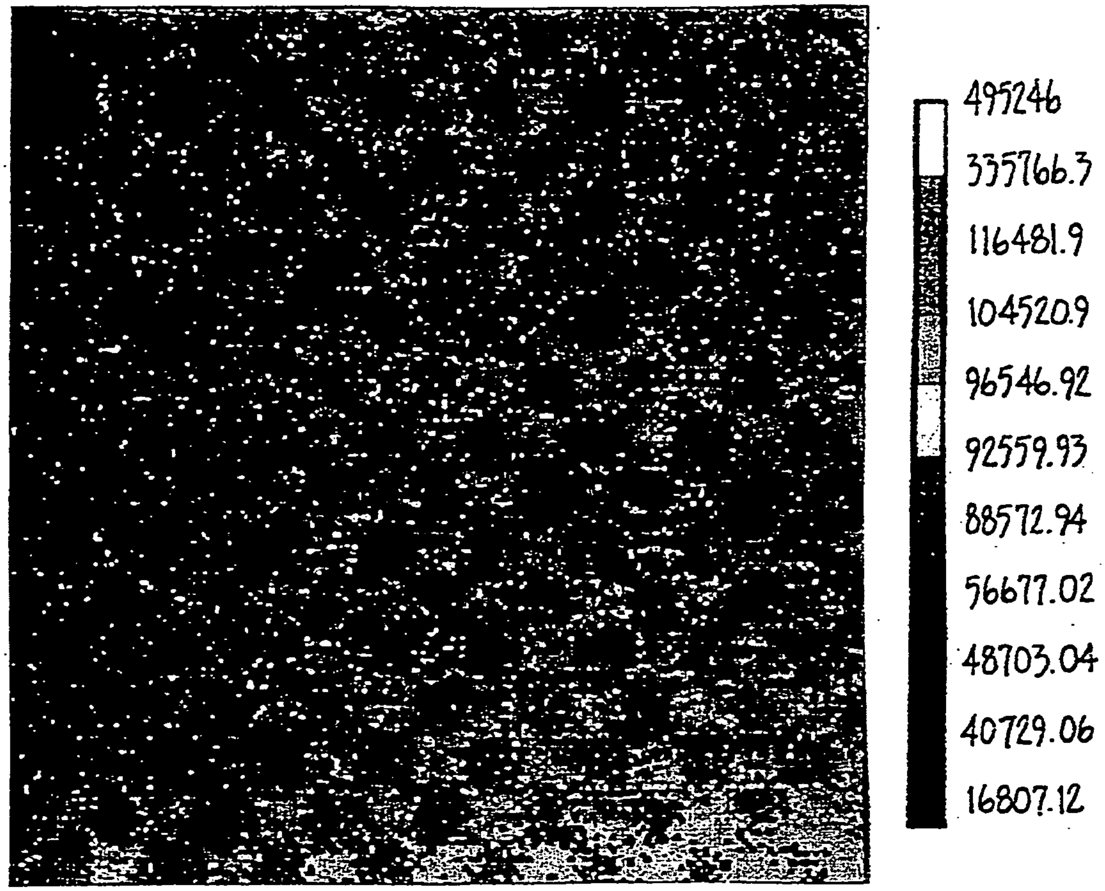
MEAN: 117723.6
VAR: 1.000047E+10
σ: 100002.3

FIG._13B.



MEAN: 104567.2
VAR: 8.025189E+09
σ: 89583.42

FIG._13C.



MEAN: 96546.92
VAR: 6.358437E+09
σ: 79739.8

FIG._13D.

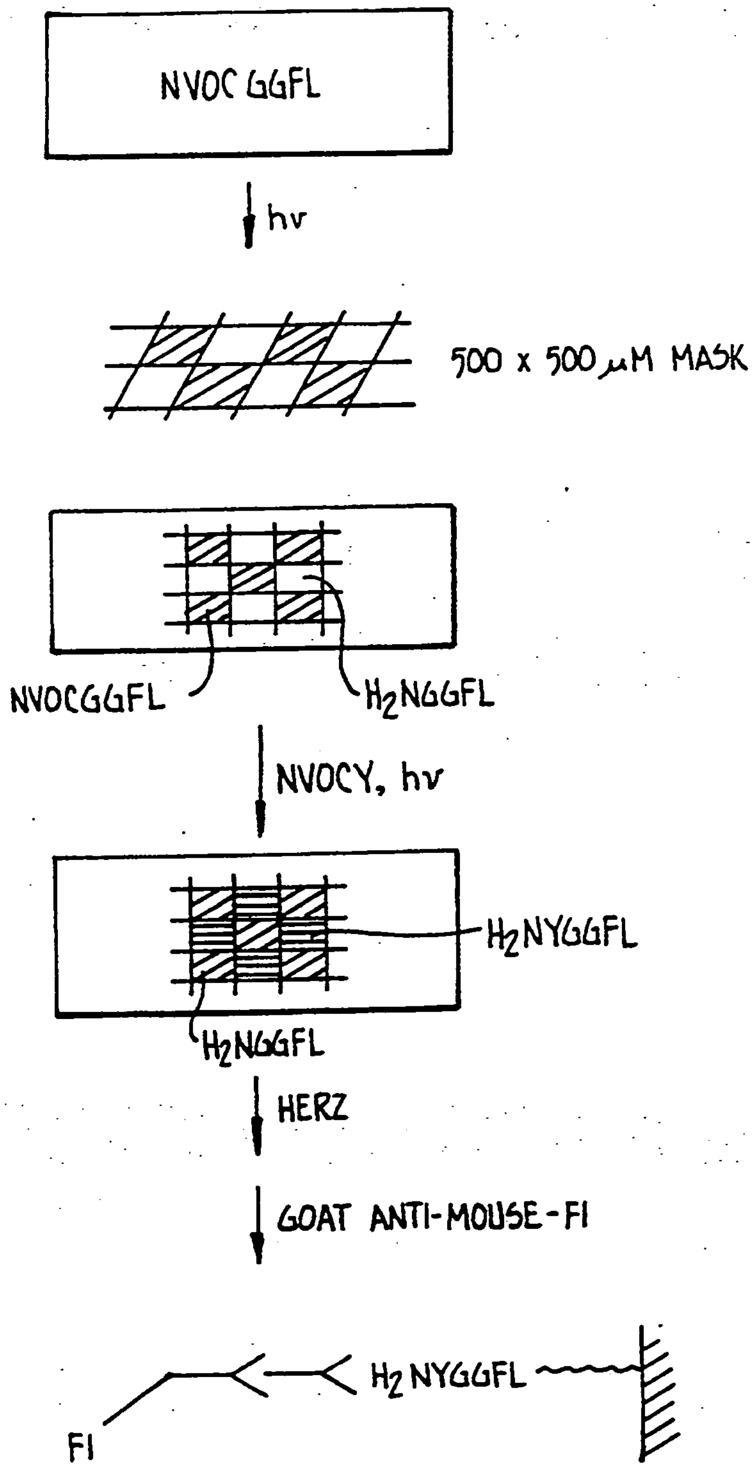

NVOC GGFL
hv
500 x 500 μM MASK
NVOCGGFL
H_2NGGFL
NVOCY, hv
$H_2NYGGFL$
H_2NGGFL
HERZ
GOAT ANTI-MOUSE-FI
$H_2NYGGFL$
FI

FIG._14B.

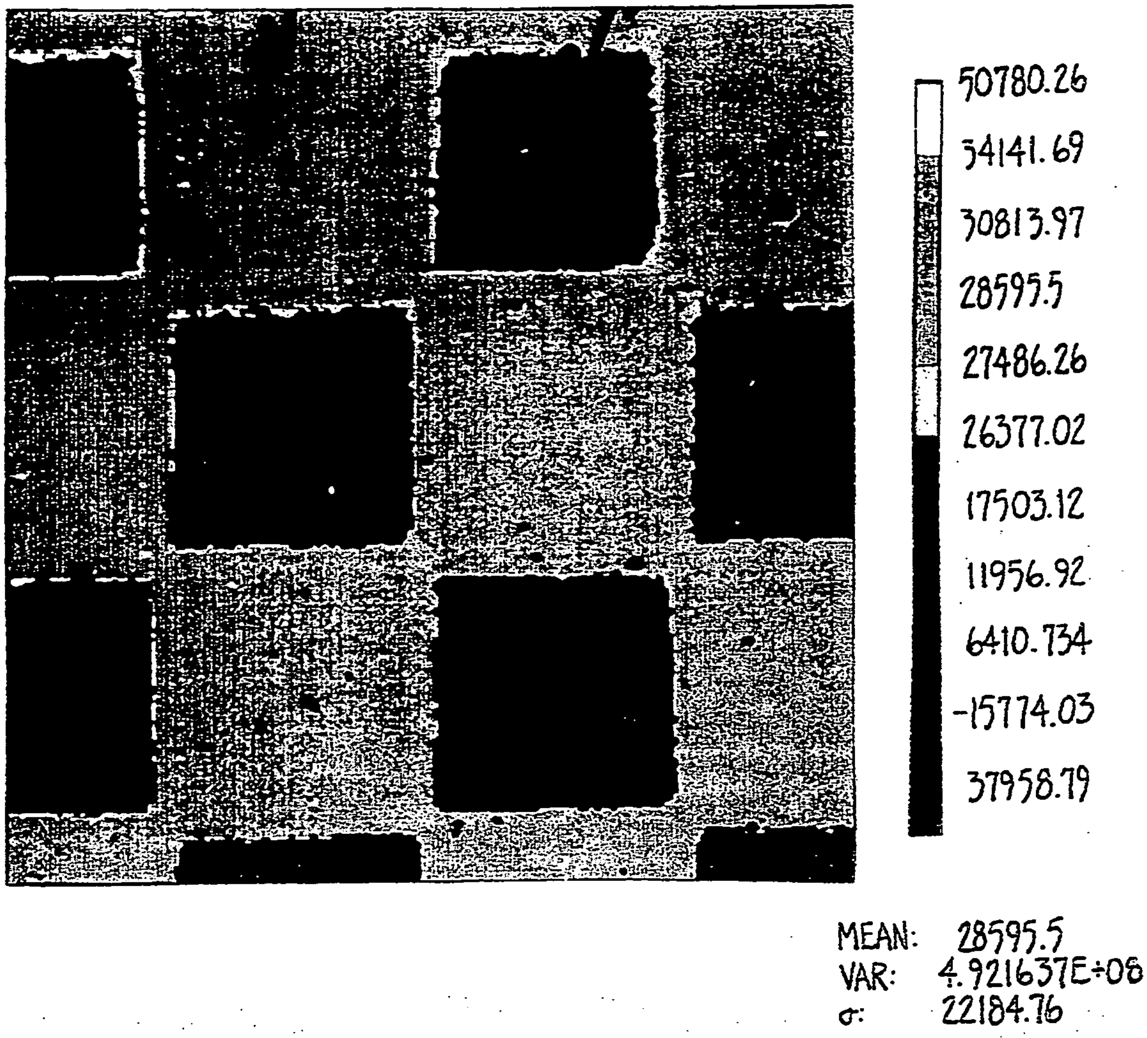
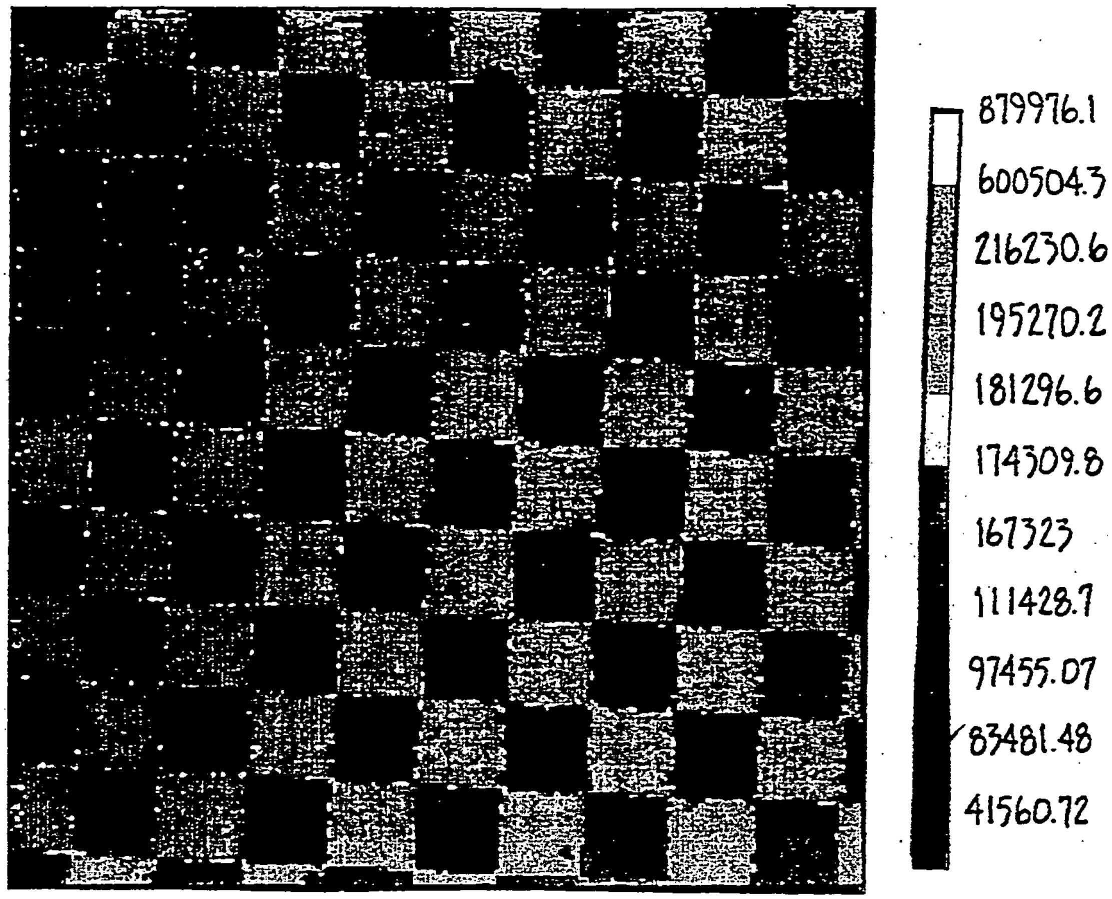
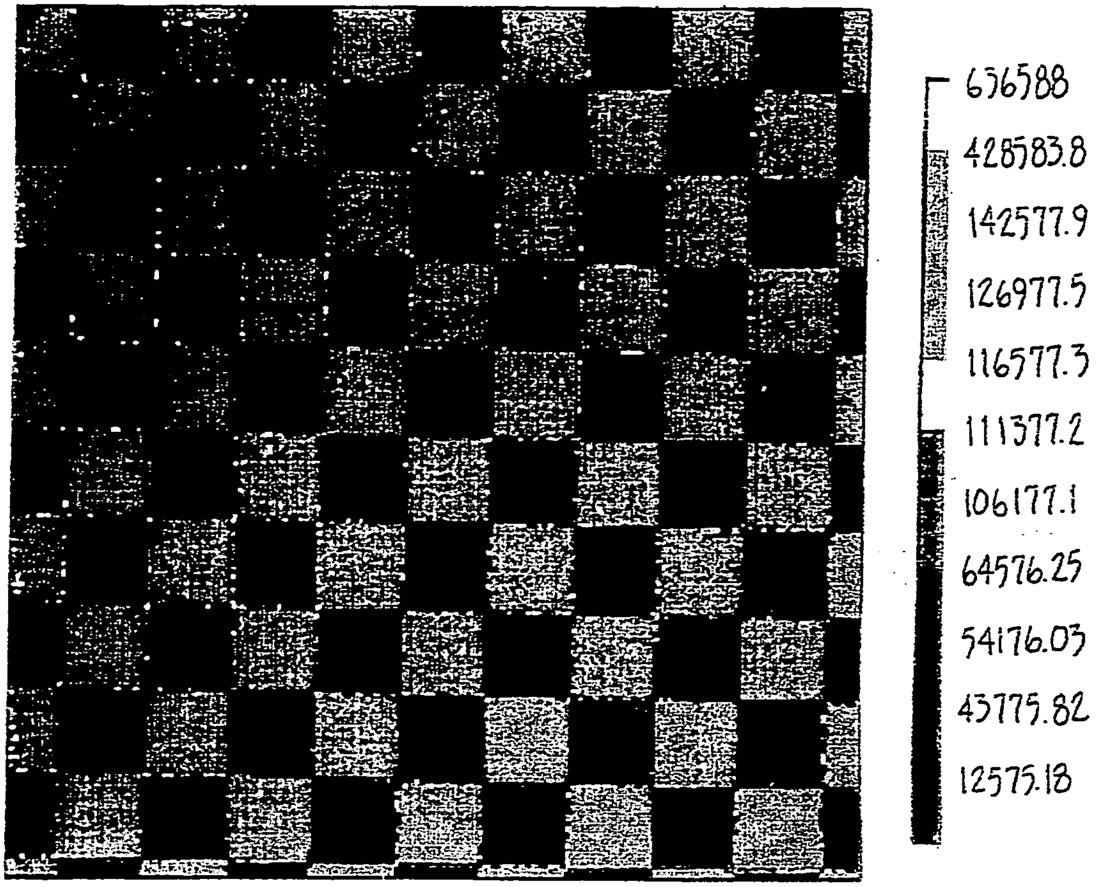

50780.26
34141.69
30813.97
28595.5
27486.26
26377.02
17503.12
11956.92
6410.734
-15774.03
37958.79
MEAN: 28595.5
VAR: 4.921637E+08
σ: 22184.76

FIG._15A.



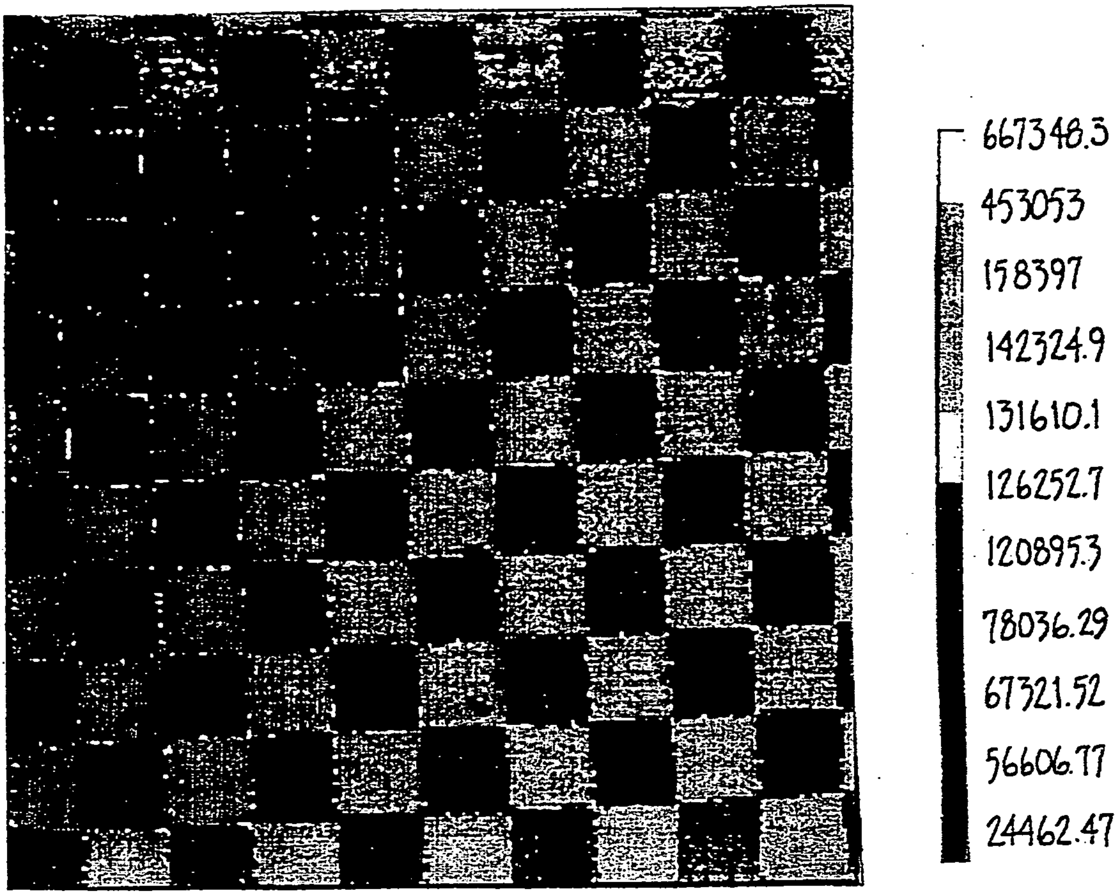
MEAN: 181296.6
VAR: 1.952612E+10
σ: 139735.9

FIG._15B.



MEAN: 116577.3
VAR: 1.081645E+10
σ: 104002.1

FIG._16.



MEAN: 131610.1
VAR: 1.148062E+10
σ: 107147.6

FIG._17.

P	A	S	G	
LPGFL	LAGFL	LSGFL	LGGFL	L
FPGFL	FAGFL	FSGFL	FGGFL	F
WPGFL	WAGFL	WSGFL	WGGFL	W
YPGFL	YAGFL	YSGFL	YGGFL	Y

L SET

FIG_18A.

p	a	s	G	
YpGFL	YaGFL	YsGFL	YGGFL	Y
fpGFL	faGFL	fsGFL	fGGFL	f
wpGFL	waGFL	wsGFL	wGGFL	w
ypGFL	yaGFL	ysGFL	yGGFL	y

D SET

FIG_18B.

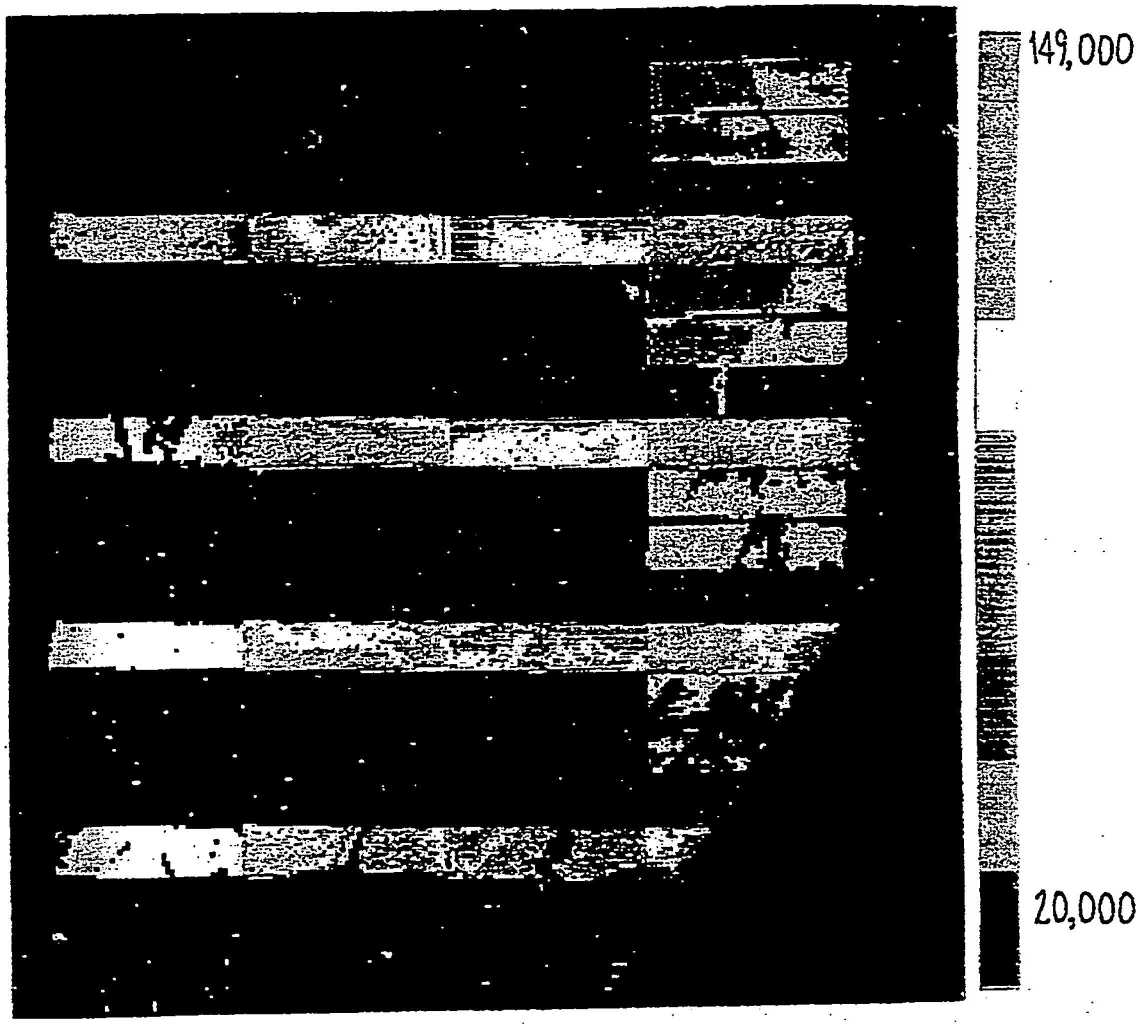


FIG._19.

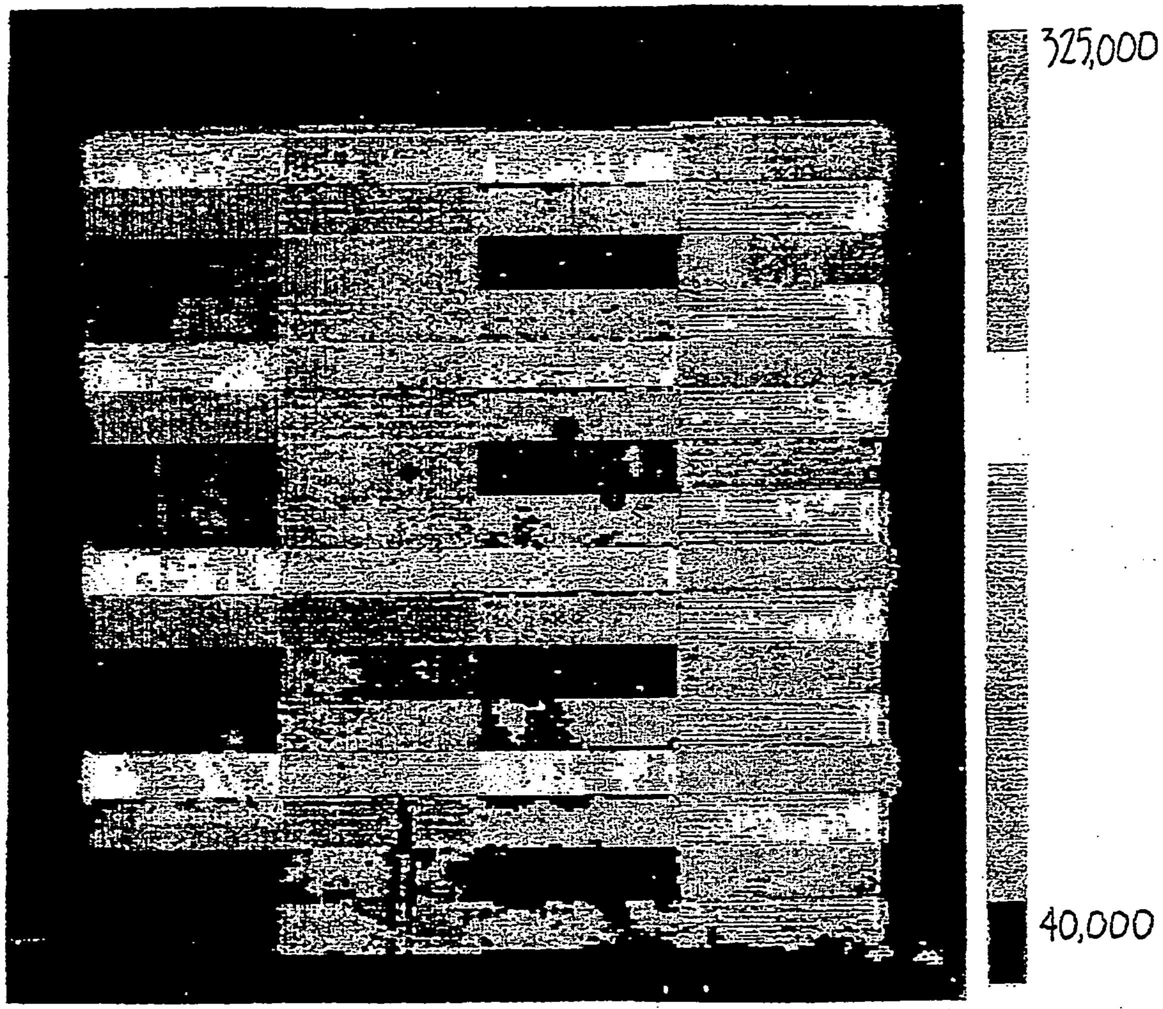


FIG._20.

ARRAY OF OLIGONUCLEOTIDES ON A SOLID SUBSTRATE

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a Rule 60 Division of U.S. application Ser. No. 850,356, filed Mar. 12, 1992, which is a Rule 60 Division of U.S. application Ser. No. 492,462, filed Mar. 7, 1990, now U.S. Pat. No. 5,143,854, which is a Continuation-in-Part of U.S. application Ser. No. 362,901, filed Jun. 7, 1989, now abandoned, all assigned to the assignee of the present invention.

The file of this patent contains drawings executed in color. Copies of this patent with color drawings will be provided by the Patent and Trademark Office upon request and payment of the necessary fee.

COPYRIGHT NOTICE

BACKGROUND OF THE INVENTION

The present inventions relate to the synthesis and placement of materials at known locations. In particular, one embodiment of the inventions provides a method and associated apparatus for preparing diverse chemical sequences at known locations on a single substrate surface. The inventions may be applied, for example, in the field of preparation of oligomer, peptide, nucleic acid, oligosaccharide, phospholipid, polymer, or drug congener preparation, especially to create sources of chemical diversity for use in screening for biological activity.

The relationship between structure and activity of molecules is a fundamental issue in the study of biological systems. Structure-activity relationships are important in understanding, for example, the function of enzymes, the ways in which cells communicate with each other, as well as cellular control and feedback systems.

Certain macromolecules are known to interact and bind to other molecules having a very specific three-dimensional spatial and electronic distribution. Any large molecule having such specificity can be considered a receptor, whether it is an enzyme catalyzing hydrolysis of a metabolic intermediate, a cell-surface protein mediating membrane transport of ions, a glycoprotein serving to identify a particular cell to its neighbors, an IgG-class antibody circulating in the plasma, an oligonucleotide sequence of DNA in the nucleus, or the like. The various molecules which receptors selectively bind are known as ligands.

Many assays are available for measuring the binding affinity of known receptors and ligands, but the information which can be gained from such experiments is often limited by the number and type of ligands which are available. Novel ligands are sometimes discovered by chance or by application of new techniques for the elucidation of molecular structure, including x-ray crystallographic analysis and recombinant genetic techniques for proteins.

Small peptides are an exemplary system for exploring the relationship between structure and function in biology. A peptide is a sequence of amino acids. When the twenty naturally occurring amino acids are condensed into polymeric molecules they form a wide variety of three-dimensional configurations, each resulting from a particular amino acid sequence and solvent condition. The number of possible pentapeptides of the 20 naturally occurring amino acids, for example, is 20^5 or 3.2 million different peptides. The likelihood that molecules of this size might be useful in receptor-binding studies is supported by epitope analysis studies showing that some antibodies recognize sequences as short as a few amino acids with high specificity. Furthermore, the average molecular weight of amino acids puts small peptides in the size range of many currently useful pharmaceutical products.

Pharmaceutical drug discovery is one type of research which relies on such a study of structure-activity relationships. In most cases, contemporary pharmaceutical research can be described as the process of discovering novel ligands with desirable patterns of specificity for biologically important receptors. Another example is research to discover new compounds for use in agriculture, such as pesticides and herbicides.

Sometimes, the solution to a rational process of designing ligands is difficult or unyielding. Prior methods of preparing large numbers of different polymers have been painstakingly slow when used at a scale sufficient to permit effective rational or random screening. For example, the "Merrifield" method (*J. Am. Chem. Soc.* (1963) 85:2149-2154, which is incorporated herein by reference for all purposes) has been used to synthesize peptides on a solid support. In the Merrifield method, an amino acid is covalently bonded to a support made of an insoluble polymer. Another amino acid with an alpha protected group is reacted with the covalently bonded amino acid to form a dipeptide. After washing, the protective group is removed and a third amino acid with an alpha protective group is added to the dipeptide. This process is continued until a peptide of a desired length and sequence is obtained. Using the Merrifield method, it is not economically practical to synthesize more than a handful of peptide sequences in a day.

To synthesize larger numbers of polymer sequences, it has also been proposed to use a series of reaction vessels for polymer synthesis. For example, a tubular reactor system may be used to synthesize a linear polymer on a solid phase support by automated sequential addition of reagents. This method still does not enable the synthesis of a sufficiently large number of polymer sequences for effective economical screening.

Methods of preparing a plurality of polymer sequences are also known in which a porous container encloses a known quantity of reactive particles, the particles being larger in size than pores of the container. The containers may be selectively reacted with desired materials to synthesize desired sequences of product molecules. As with other methods known in the art, this method cannot practically be used to synthesize a sufficient variety of polypeptides for effective screening.

Other techniques have also been described. These methods include the synthesis of peptides on 96 plastic pins which fit the format of standard microtiter plates. Unfortunately, while these techniques have been somewhat useful, substantial problems remain. For example, these methods continue to be limited in the diversity of sequences which can be economically synthesized and screened.

From the above, it is seen that an improved method and apparatus for synthesizing a variety of chemical sequences at known locations is desired.

SUMMARY OF THE INVENTION

An improved method and apparatus for the preparation of a variety of polymers is disclosed.

In one preferred embodiment, linker molecules are provided on a substrate. A terminal end of the linker molecules is provided with a reactive functional group protected with a photoremovable protective group. Using lithographic methods, the photoremovable protective group is exposed to light and removed from the linker molecules in first selected regions. The substrate is then washed or otherwise contacted with a first monomer that reacts with exposed functional groups on the linker molecules. In a preferred embodiment, the monomer is an amino acid containing a photoremovable protective group at its amino or carboxy terminus and the linker molecule terminates in an amino or carboxy acid group bearing a photoremovable protective group.

A second set of selected regions is, thereafter, exposed to light and the photoremovable protective group on the linker molecule/protected amino acid is removed at the second set of regions. The substrate is then contacted with a second monomer containing a photoremovable protective group for reaction with exposed functional groups. This process is repeated to selectively apply monomers until polymers of a desired length and desired chemical sequence are obtained. Photolabile groups are then optionally removed and the sequence is, thereafter, optionally capped. Side chain protective groups, if present, are also removed.

By using the lithographic techniques disclosed herein, it is possible to direct light to relatively small and precisely known locations on the substrate. It is, therefore, possible to synthesize polymers of a known chemical sequence at known locations on the substrate.

The resulting substrate will have a variety of uses including, for example, screening large numbers of polymers for biological activity. To screen for biological activity, the substrate is exposed to one or more receptors such as antibodies whole cells, receptors on vesicles, lipids, or any one of a variety of other receptors. The receptors are preferably labeled with, for example, a fluorescent marker, radioactive marker, or a labeled antibody reactive with the receptor. The location of the marker on the substrate is detected with, for example, photon detection or autoradiographic techniques. Through knowledge of the sequence of the material at the location where binding is detected, it is possible to quickly determine which sequence binds with the receptor and, therefore, the technique can be used to screen large numbers of peptides. Other possible applications of the inventions herein include diagnostics in which various antibodies for particular receptors would be placed on a substrate and, for example, blood sera would be screened for immune deficiencies. Still further applications include, for example, selective "doping" of organic materials in semiconductor devices, and the like.

In connection with one aspect of the invention an improved reactor system for synthesizing polymers is also disclosed. The reactor system includes a substrate mount which engages a substrate around a periphery thereof. The substrate mount provides for a reactor space between the substrate and the mount through or into which reaction fluids are pumped or flowed. A mask is placed on or focused on the substrate and illuminated so as to deprotect selected regions of the substrate in the reactor space. A monomer is pumped through the reactor space or otherwise contacted with the substrate and reacts with the deprotected regions. By selectively deprotecting regions on the substrate and flowing predetermined monomers through the reactor space, desired polymers at known locations may be synthesized.

Improved detection apparatus and methods are also disclosed. The detection method and apparatus utilize a substrate having a large variety of polymer sequences at known locations on a surface thereof. The substrate is exposed to a fluorescently labeled receptor which binds to one or more of the polymer sequences. The substrate is placed in a microscope detection apparatus for identification of locations where binding takes place. The microscope detection apparatus includes a monochromatic or polychromatic light source for directing light at the substrate, means for detecting fluoresced light from the substrate, and means for determining a location of the fluoresced light. The means for detecting light fluoresced on the substrate may in some embodiments include a photon counter. The means for determining a location of the fluoresced light may include an x/y translation table for the substrate. Translation of the slide and data collection are recorded and managed by an appropriately programmed digital computer.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1 illustrates masking and irradiation of a substrate at a first location. The substrate is shown in cross-section;

FIG. 2 illustrates the substrate after application of a monomer "A";

FIG. 3 illustrates irradiation of the substrate at a second location;

FIG. 4 illustrates the substrate after application of monomer "B";

FIG. 5 illustrates irradiation of the "A" monomer;

FIG. 6 illustrates the substrate after a second application of "B";

FIG. 7 illustrates a completed substrate;

FIGS. 8A and 8B illustrate alternative embodiments of a reactor system for forming a plurality of polymers on a substrate;

FIG. 9 illustrates a detection apparatus for locating fluorescent markers on the substrate;

FIGS. 10A–10M illustrate the method as it is applied to the production of the trimers of monomers "A" and "B";

FIGS. 11A and 11B are fluorescence traces for standard fluorescent beads;

FIGS. 12A and 12B are fluorescence curves for NVOC (6-nitroveratryloxycarbonyl) slides not exposed and exposed to light respectively;

FIGS. 13A to 13D are fluorescence plots of slides exposed through 100 μm, 50 μm, 20 μm, and 10 μm masks; 14A and 14B illustrate formation of YGGFL (a peptide of sequence H2N-tyrosine-glycine-glycine-phenylalanine-leucine-CO_2H) and GGFL (a peptide of sequence H_2N-glycine-glycine-phenylalanine-leucine-CO_2H), followed by exposure to labeled Herz antibody (an antibody that recognizes YGGFL but not GGFL);

FIGS. 15A and 15B fluorescence plots of a slide with a checkerboard pattern of YGGFL and GGFL exposed to labeled Herz antibody; FIG. 15A illustrates a 500×500 μm mask which has been focused on the substrate according to FIG. 8A while FIG. 15B illustrates a 50×50 μm mask placed in direct contact with the substrate in accord with FIG. 8B;

FIG. 16 is a fluorescence plot of YGGFL and PGGFL synthesized in a 50 μm checkerboard pattern;

FIG. 17 is a fluorescence plot of YPGGFL and YGGFL synthesized in a 50 μm checkerboard pattern;

FIGS. 18A and 18B illustrate the mapping of sixteen sequences synthesized on two different glass slides;

FIG. 19 is a fluorescence plot of the slide illustrated in FIG. 18A; and

FIG. 20 is a fluorescence plot of the slide illustrated in FIG. 10B.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

CONTENTS

I. Glossary

The following terms are intended to have the following general meanings as they are used herein:

1. Complementary: Refers to the topological compatibility or matching together of interacting surfaces of a ligand molecule and its receptor. Thus, the receptor and its ligand can be described as complementary, and furthermore, the contact surface characteristics are complementary to each other.
2. Epitope: The portion of an antigen molecule which is delineated by the area of interaction with the subclass of receptors known as antibodies.
3. Ligand: A ligand is a molecule that is recognized by a particular receptor. Examples of ligands that can be investigated by this invention include, but are not restricted to, agonists and antagonists for cell membrane receptors, toxins and venoms, viral epitopes, hormones (e.g., steroids, etc.), hormone receptors, peptides, enzymes, enzyme substrates, cofactors, drugs (e.g., opiates, etc), lectins, sugars, oligonucleotides, nucleic acids, oligosaccharides, proteins, and monoclonal antibodies.
4. Monomer: A member of the set of small molecules which can be joined together to form a polymer. The set of monomers includes but is not restricted to, for example, the set of common L-amino acids, the set of D-amino acids, the set of synthetic amino acids, the set of nucleotides and the set of pentoses and hexoses. As used herein, monomers refers to any member of a basis set for synthesis of a polymer. For example, dimers of L-amino acids form a basis set of 400 monomers for synthesis of polypeptides. Different basis sets of monomers may be used at successive steps in the synthesis of a polymer.
5. Peptide: A polymer in which the monomers are alpha amino acids and which are joined together through amide bonds and alternatively referred to as a polypeptide. In the context of this specification it should be appreciated that the amino acids may be the L-optical isomer or the D-optical isomer. Peptides are more than two amino acid monomers long, and often more than 20 amino acid monomers long. Standard abbreviations for amino acids are used (e.g., P for proline). These abbreviations are included in Stryer, *Biochemstry*, Third Ed., 1988, which is incorporated herein by reference for all purposes.
6. Radiation: Energy which may be selectively applied including energy having a wavelength of between 10^{-14} and 10^4 meters including, for example, electron beam radiation, gamma radiation, x-ray radiation, ultraviolet radiation, visible light, infrared radiation, microwave radiation, and radio waves. "Irradiation" refers to the application of radiation to a surface.
7. Receptor: A molecule that has an affinity for a given ligand. Receptors may be naturally-occuring or manmade molecules. Also, they can be employed in their unaltered state or as aggregates with other species. Receptors may be attached, covalently or noncovalently, to a binding member, either directly or via a specific binding substance. Examples of receptors which can be employed by this invention include, but are not restricted to, antibodies, cell membrane receptors, monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells or other materials), drugs, polynucleotides, nucleic acids, peptides, cofactors, lectins, sugars, polysaccharides, cells, cellular membranes, and organelles. Receptors are sometimes referred to in the art as anti-ligands. As the term receptors is used herein, no difference in meaning is intended. A "Ligand Receptor Pair" is formed when two macromolecules have combined through molecular recognition to form a complex.

Other examples of receptors which can be investigated by this invention include but are not restricted to:

a) Microorganism receptors: Determination of ligands which bind to receptors, such as specific transport proteins or enzymes essential to survival of microorganisms, is useful in a new class of antibiotics. Of particular value would be antibiotics against opportunistic fungi, protozoa, and those bacteria resistant to the antibiotics in current use.

b) Enzymes: For instance, the binding site of enzymes such as the enzymes responsible for cleaving neurotransmitters; determination of ligands which bind to certain receptors to modulate the action of the enzymes which cleave the different neurotransmitters is useful in the development of drugs which can be used in the treatment of disorders of neurotransmission.

c) Antibodies: For instance, the invention may be useful in investigating the ligand-binding site on the antibody molecule which combines with the epitope of an antigen of interest; determining a sequence that mimics an antigenic epitope may lead to the development of vaccines of which the immunogen is based on one or more of such sequences or lead to the development of related diagnostic agents or compounds useful in therapeutic treatments such as for auto immune diseases (e.g., by blocking the binding of the "self" antibodies).

d) Nucleic Acids: Sequences of nucleic acids may be synthesized to establish DNA or RNA binding sequences.

e) Catalytic Polypeptides: Polymers, preferably polypeptides, which are capable of promoting a chemical reaction involving the conversion of one or more reactants to one or more products. Such polypeptides generally include a binding site specific for at least one reactant or reaction intermediate and an active functionality proximate to the binding site, which functionality is capable of chemically modifying the bound reactant. Catalytic polypeptides are described in, for example, U.S. application Ser. No. 404,920, which is incorporated herein by reference for all purposes.

f) Hormone receptors: For instance, the receptors for insulin and growth hormone. Determination of the ligands which bind with high affinity to a receptor is useful in the development of, for example, an oral replacement of the daily injections which diabetics must take to relieve the symptoms of diabetes, and in the other case, a replacement for the scarce human growth hormone which can only be obtained from cadavers or by recombinant DNA technology. Other examples are the vasoconstrictive hormone receptors; determination of those ligands which bind to a receptor may lead to the development of drugs to control blood pressure.

g) Opiate receptors: Determination of ligands which bind to the opiate receptors in the brain is useful in the development of less-addictive replacements for morphine and related drugs.

8. Substrate: A material having a rigid or semi-rigid surface. In many embodiments, at least one surface of the substrate will be substantially flat, although in some embodiments it may be desirable to physically separate synthesis regions for different polymers with, for example, wells, raised regions, etched trenches, or the like. According to other embodiments, small beads may be provided on the surface which may be released upon completion of the synthesis.

9. Protective Group: A material which is bound to a monomer unit and which may be spatially removed upon selective exposure to an activator such as electromagnetic radiation. Examples of protective groups with utility herein include Nitroveratryloxy carbonyl, Nitrobenzyloxy carbonyl, Dimethyl dimethoxybenzyloxy carbonyl, 5-Bromo-7-nitroindolinyl, o-Hydroxy-α-methyl cinnamoyl, and 2-Oxymethylene anthraquinone. Other examples of activators include ion beams, electric fields, magnetic fields, electron beams, x-ray, and the like.

10. Predefined Region: A predefined region is a localized area on a surface which is, was, or is intended to be activated for formation of a polymer. The predefined region may have any convenient shape, e.g., circular, rectangular, elliptical, wedge-shaped, etc. For the sake of brevity herein, "predefined regions" are sometimes referred to simply as "regions."

11. Substantially Pure: A polymer is considered to be "substantially pure" within a predefined region of a substrate when it exhibits characteristics that distinguish it from other predefined regions. Typically, purity will be measured in terms of biological activity or function as a result of uniform sequence. Such characteristics will typically be measured by way of binding with a selected ligand or receptor.

II. General

The present invention provides methods and apparatus for the preparation and use of a substrate having a plurality of polymer sequences in predefined regions. The invention is described herein primarily with regard to the preparation of molecules containing sequences of amino acids, but could readily be applied in the preparation of other polymers. Such polymers include, for example, both linear and cyclic polymers of nucleic acids, polysaccharides, phospholipids, and peptides having either α-, β-, or ω-amino acids, heteropolymers in which a known drug is covalently bound to any of the above, polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneimines, polyarylene sulfides, polysiloxanes, polyimides, polyacetates, or other polymers which will be apparent upon review of this disclosure. In a preferred embodiment, the invention herein is used in the synthesis of peptides.

The prepared substrate may, for example, be used in screening a variety of polymers as ligands for binding with a receptor, although it will be apparent that the invention could be used for the synthesis of a receptor for binding with a ligand. The substrate disclosed herein will have a wide variety of other uses. Merely by way of example, the invention herein can be used in determining peptide and nucleic acid sequences which bind to proteins, finding sequence-specific binding drugs, identifying epitopes recognized by antibodies, and evaluation of a variety of drugs for clinical and diagnostic applications, as well as combinations of the above.

The invention preferably provides for the use of a substrate "S" with a surface. Linker molecules "L" are optionally provided on a surface of the substrate. The purpose of the linker molecules, in some embodiments, is to facilitate receptor recognition of the synthesized polymers.

Optionally, the linker molecules may be chemically protected for storage purposes. A chemical storage protective group such as t-BOC (t-butoxycarbonyl) may be used in some embodiments. Such chemical protective groups would be chemically removed upon exposure to, for example, acidic solution and would serve to protect the surface during storage and be removed prior to polymer preparation.

On the substrate or a distal end of the linker molecules, a functional group with a protective group P_0 is provided. The protective group P_0 may be removed upon exposure to radiation, electric fields, electric currents, or other activators to expose the functional group.

In a preferred embodiment, the radiation is ultraviolet (UV), infrared (IR), or visible light. As more fully described below, the protective group may alternatively be an electrochemically-sensitive group which may be removed in the presence of an electric field. In still further alternative embodiments, ion beams, electron beams, or the like may be used for deprotection.

In some embodiments, the exposed regions and, therefore, the area upon which each distinct polymer sequence is synthesized are smaller than about 1 cm^2 or less than 1 mm^2. In preferred embodiments the exposed area is less than about 10,000 μm^2 or, more preferably, less than 100 μm^2 and may, in some embodiments, encompass the binding site for as few as a single molecule. Within these regions, each polymer is preferably synthesized in a substantially pure form.

Concurrently or after exposure of a known region of the substrate to light, the surface is contacted with a first monomer unit M_1 which reacts with the functional group which has been exposed by the deprotection step. The first monomer includes a protective group P_1. P_1 may or may not be the same as P_0.

Accordingly, after a first cycle, known first regions of the surface may comprise the sequence:

$$S\text{-}L\text{-}M_1\text{-}P_1$$

while remaining regions of the surface comprise the sequence:

$$S\text{-}L\text{-}P_0.$$

Thereafter, second regions of the surface (which may include the first region) are exposed to light and contacted with a second monomer M_2 (which may or may not be the same as M_1) having a protective group P_2. P_2 may or may not be the same as P_0 and P_1. After this second cycle, different regions of the substrate may comprise one or more of the following sequences:

$$S\text{-}L\text{-}M_1\text{-}M_2\text{-}P_2\ S\text{-}L\text{-}M_2\text{-}P_2\ S\text{-}L\text{-}M_1\text{-}P_1 \text{ and/or } S\text{-}L\text{-}P_0.$$

The above process is repeated until the substrate includes desired polymers of desired lengths. By controlling the locations of the substrate exposed to light and the reagents exposed to the substrate following exposure, the location of each sequence will be known.

Thereafter, the protective groups are removed from some or all of the substrate and the sequences are, optionally, capped with a capping unit C. The process results in a substrate having a surface with a plurality of polymers of the following general formula:

$$S\text{-}[L]\text{-}(M_i)\text{-}(M_j)\text{-}(M_k)\ldots(M_x)\text{-}[C]$$

where square brackets indicate optional groups, and $M_i \ldots M_x$ indicates any sequence of monomers. The number of monomers could cover a wide variety of values, but in a preferred embodiment they will range from 2 to 100.

In some embodiments a plurality of locations on the substrate polymers are to contain a common monomer subsequence. For example, it may be desired to synthesize a sequence $S\text{-}M_1\text{-}M_2\text{-}M_3$ at first locations and a sequence $S\text{-}M_4\text{-}M_2\text{-}M_3$ at second locations. The process would commence with irradiation of the first locations followed by contacting with $M_1\text{-}P$, resulting in the sequence $S\text{-}M_1\text{-}P$ at the first location. The second locations would then be irradiated and contacted with $M_4\text{-}P$, resulting in the sequence $S\text{-}M_4\text{-}P$ at the second locations. Thereafter both the first and second locations would be irradiated and contacted with the dimer $M_2\text{-}M_3$, resulting in the sequence $S\text{-}M_1\text{-}M_2\text{-}M_3$ at the first locations and $S\text{-}M_4\text{-}M_2\text{-}M_3$ at the second locations. Of course, common subsequences of any length could be utilized including those in a range of 2 or more monomers, 2 to 100 monomers, 2 to 20 monomers, and a most preferred range of 2 to 3 monomers.

According to other embodiments, a set of masks is used for the first monomer layer and, thereafter, varied light wavelengths are used for selective deprotection. For example, in the process discussed above, first regions are first exposed through a mask and reacted with a first monomer having a first protective group P_1, which is removable upon exposure to a first wavelength of light (e.g., IR). Second regions are masked and reacted with a second monomer having a second protecive group P_2, which is removable upon exposure to a second wavelength of light (e.g., UV). Thereafter, masks become unnecessary in the synthesis because the entire substrate may be exposed alternatively to the first and second wavelengths of light in the deprotection cycle.

The polymers prepared on a substrate according to the above methods will have a variety of uses including, for example, screening for biological activity. In such screening activities, the substrate containing the sequences is exposed to an unlabeled or labeled receptor such as an antibody, receptor on a cell, phospholipid vesicle, or any one of a variety of other receptors. In one preferred embodiment the polymers are exposed to a first, unlabeled receptor of interest and, thereafter, exposed to a labeled receptor-specific recognition element, which is, for example, an antibody. This process will provide signal amplification in the detection stage.

The receptor molecules may bind with one or more polymers on the substrate. The presence of the labeled receptor and, therefore, the presence of a sequence which binds with the receptor is detected in a preferred embodiment through the use of autoradiography, detection of fluorescence with a charge-coupled device, fluorescence microscopy, or the like. The sequence of the polymer at the locations where the receptor binding is detected may be used to determine all or part of a sequence which is complementary to the receptor.

Use of the invention herein is illustrated primarily with reference to screening for biological activity. The invention will, however, find many other uses. For example, the invention may be used in information storage (e.g., on optical disks), production of molecular electronic devices, production of stationary phases in separation sciences, production of dyes and brightening agents, photography, and in immobilization of cells, proteins, lectins, nucleic acids, polysaccharides and the like in patterns on a surface via molecular recognition of specific polymer sequences. By synthesizing the same compound in adjacent, progressively differing concentrations, a gradient will be established to control chemotaxis or to develop diagnostic dipsticks which, for example, titrate an antibody against an increasing amount of antigen. By synthesizing several catalyst molecules in close proximity, more efficient multistep conversions may be achieved by "coordinate immobilization." Coordinate immobilization also may be used for electron transfer systems, as well as to provide both structural integrity and other desirable properties to materials such as lubrication, wetting, etc.

According to alternative embodiments, molecular biodistribution or pharmacokinetic properties may be examined. For example, to assess resistance to intestinal or serum proteases, polymers may be capped with a fluorescent tag and exposed to biological fluids of interest.

III. Polymer Synthesis

FIG. 1 illustrates one embodiment of the invention disclosed herein in which a substrate 2 is shown in cross-section. Essentially, any conceivable substrate may be employed in the invention. The substrate may be biological, nonbiological, organic, inorganic, or a combination of any of these, existing as particles, strands, precipitates, gels, sheets, tubing, spheres, containers, capillaries, pads, slices, films, plates, slides, etc. The substrate may have any convenient shape, such as a disc, square, sphere, circle, etc. The substrate is preferably flat but may take on a variety of alternative surface configurations. For example, the substrate may contain raised or depressed regions on which the synthesis takes place. The substrate and its surface preferably form a rigid support on which to carry out the reactions described herein. The substrate and its surface is also chosen to provide appropriate light-absorbing characteristics. For instance, the substrate may be a polymerized Langmuir Blodgett film, functionalized glass, Si, Ge, GaAs, GaP, SiO_2, SIN_4, modified silicon, or any one of a wide variety of gels or polymers such as (poly)-tetrafluoroethylene, (poly)vinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof. Other substrate materials will be readily apparent to those of skill in the art upon review of this disclosure. In a preferred embodiment the substrate is flat glass or single-crystal silicon with surface relief features of less than 10 Å.

According to some embodiments, the surface of the substrate is etched using well known techniques to provide for desired surface features. For example, by way of the formation of trenches, v-grooves, mesa structures, or the like, the synthesis regions may be more closely placed within the focus point of impinging light, be provided with reflective "mirror" structures for maximization of light collection from fluorescent sources, or the like.

Surfaces on the solid substrate will usually, though not always, be composed of the same material as the substrate. Thus, the surface may be composed of any of a wide variety of materials, for example, polymers, plastics, resins, polysaccharides, silica or silica-based materials, carbon, metals, inorganic glasses, membranes, or any of the above-listed substrate materials. In some embodiments the surface may provide for the use of caged binding members which are attached firmly to the surface of the substrate. Preferably, the surface will contain reactive groups, which could be carboxyl, amino, hydroxyl, or the like. Most preferably, the surface will be optically transparent and will have surface Si—OH functionalities, such as are found on silica surfaces.

The surface 4 of the substrate is preferably provided with a layer of linker molecules 6, although it will be understood that the linker molecules are not required elements of the invention. The linker molecules are preferably of sufficient length to permit polymers in a completed substrate to interact freely with molecules exposed to the substrate. The linker molecules should be 6–50 atoms long to provide sufficient exposure. The linker molecules may be, for example, aryl acetylene, ethylene glycol oligomers containing 2–10 monomer units, diamines, diacids, amino acids, or combinations thereof. Other linker molecules may be used in light of this disclsoure.

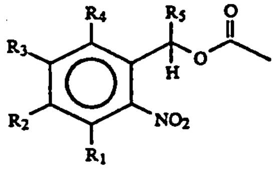
According to alternative embodiments, the linker molecules are selected based upon their hydrophilic/-hydrophobic properties to improve presentation of synthesized polymers to certain receptors. For example, in the case of a hydrophilic receptor, hydrophilic linker molecules will be preferred so as to permit the receptor to more closely approach the synthesized polymer.

According to another alternative embodiment, linker molecules are also provided with a photocleavable group at an intermediate position. The photocleavable group is preferably cleavable at a wavelength different from the protective group. This enables removal of the various polymers following completion of the synthesis by way of exposure to the different wavelengths of light.

The linker molecules can be attached to the substrate via carbon-carbon bonds using, for example, (poly)trifluorochloroethylene surfaces, or preferably, by siloxane bonds (using, for example, glass or silicon oxide surfaces). Siloxane bonds with the surface of the substrate may be formed in one embodiment via reactions of linker molecules bearing trichlorosilyl groups. The linker molecules may optionally be attached in an ordered array, i.e., as parts of the head groups in a polymerized Langmuir Blodgett film. In alternative embodiments, the linker molecules are adsorbed to the surface of the substrate.

The linker molecules and monomers used herein are provided with a functional group to which is bound a protective group. Preferably, the protective group is on the distal or terminal end of the linker molecule opposite the substrate. The protective group may be either a negative protective group (i.e., the protective group renders the linker molecules less reactive with a monomer upon exposure) or a positive protective group (i.e., the protective group renders the linker molecules more reactive with a monomer upon exposure). In the case of negative protective groups an additional step of reactivation will be required. In some embodiments, this will be done by heating.

The protective group on the linker molecules may be selected from a wide variety of positive light-reactive groups preferably including nitro aromatic compounds such as o-nitrobenzyl derivatives or benzylsulfonyl. In a preferred embodiment, 6-nitroveratryloxycarbonyl (NVOC), 2-nitrobenzyloxycarbonyl (NBOC) or α,α-dimethyl-dimethoxybenzyloxycarbonyl (DDZ) is used. In one embodiment, a nitro aromatic compound containing a benzylic hydrogen ortho to the nitro group is used, i.e., a chemical of the form:



where R_1 is alkoxy, alkyl, halo, aryl, alkenyl, or hydrogen; R_2 is alkoxy, alkyl, halo, aryl, nitro, or hydrogen; R_3 is alkoxy, alkyl, halo, nitro, aryl, or hydrogen; R_4 is alkoxy, alkyl, hydrogen, aryl, halo, or nitro; and R_5 is alkyl, alkynyl, cyano, alkoxy, hydrogen, halo, aryl, or alkenyl. Other materials which may be used include o-hydroxy-α-methyl cinnamoyl derivatives. Photoremovable protective groups are described in, for example, Patchornik, *J. Am. Chem. Soc.* (1970) 92:6333 and Amit et al., *J. Org. Chem.* (1974) 39:192, both of which are incorporated herein by reference.

In an alternative embodiment the positive reactive group is activated for reaction with reagents in solution. For example, a 5-bromo-7-nitro indoline group, when bound to a carbonyl, undergoes reaction upon exposure to light at 420 nm.

In a second alternative embodiment, the reactive group on the linker molecule is selected from a wide variety of negative light-reactive groups including a cinammate group.

Alternatively, the reactive group is activated or deactivated by electron beam lithography, x-ray lithography, or any other radiation. Suitable reactive groups for electron beam lithography include sulfonyl. Other methods may be used including, for example, exposure to a current source. Other reactive groups and methods of activation may be used in light of this disclosure.

As shown in FIG. 1, the linking molecules are preferably exposed to, for example, light through a suitable mask 8 using photolithographic techniques of the type known in the semiconductor industry and described in, for example, Sze, *VLSI Technology*, McGraw-Hill (1983), and Mead et al., *Introduction to VLSI Systems*, Addison-Wesley (1980), which are incorporated herein by reference for all purposes. The light may be directed at either the surface containing the protective groups or at the back of the substrate, so long as the substrate is transparent to the wavelength of light needed for removal of the protective groups. In the embodiment shown in FIG. 1, light is directed at the surface of the substrate containing the protective groups. FIG. 1 illustrates the use of such masking techniques as they are applied to a positive reactive group so as to activate linking molecules and expose functional groups in areas 10*a* and 10*b*.

The mask 8 is in one embodiment a transparent support material selectively coated with a layer of opaque material. Portions of the opaque material are removed, leaving opaque material in the precise pattern desired on the substrate surface. The mask is brought into close proximity with, imaged on, or brought directly into contact with the substrate surface as shown in FIG. 1. "Openings" in the mask correspond to locations on the substrate where it is desired to remove photoremovable protective groups from the substrate. Alignment may be performed using conventional alignment techniques in which alignment marks (not shown) are used to accurately overlay successive masks with previous patterning steps, or more sophisticated techniques may be used. For example, interferometric techniques such as the one described in Flanders et al., "A New Interferometric Alignment Technique," *App. Phys. Lett.* (1977) 31:426–428, which is incorporated herein by reference, may be used.

To enhance contrast of light applied to the substrate, it is desirable to provide contrast enhancement materials between the mask and the substrate according to some embodiments. This contrast enhancement layer may comprise a molecule which is decomposed by light such as quinone diazide or a material which is transiently bleached at the wavelength of interest. Transient bleaching of materials will allow greater penetration where light is applied, thereby enhancing contrast. Alternatively, contrast enhancement may be provided by way of a cladded fiber optic bundle.

The light may be from a conventional incandescent source, a laser, a laser diode, or the like. If non-collimated sources of light are used it may be desirable to provide a thick- or multi-layered mask to prevent spreading of the light onto the substrate. It may, further, be desirable in some embodiments to utilize groups which are sensitive to different wavelengths to control synthesis. For example, by using groups which are sensitive to different wavelengths, it is possible to select branch positions in the synthesis of a polymer or eliminate certain masking steps. Several reactive groups along with their corresponding wavelengths for deprotection are provided in Table 1.

TABLE 1

Group	Approximate Deprotection Wavelength
Nitroveratryloxy carbonyl (NVOC)	UV (300–400 nm)
Nitrobenzyloxy carbonyl (NBOC)	UV (300–350 nm)
Dimethyl dimethoxybenzyloxy carbonyl	UV (280–300 nm)
5-Bromo-7-nitroindolinyl	UV (420 nm)
o-Hydroxy-α-methyl cinnamoyl	UV (300–350 nm)
2-Oxymethylene anthraquinone	UV (350 nm)

While the invention is illustrated primarily herein by way of the use of a mask to illuminate selected regions the substrate, other techniques may also be used. For example, the substrate may be translated under a modulated laser or diode light source. Such techniques are discussed in, for example, U.S. Pat. No. 4,719,615 (Feyrer et al.), which is incorporated herein by reference. In alternative embodiments a laser galvanometric scanner is utilized. In other embodiments, the synthesis may take place on or in contact with a conventional liquid crystal (referred to herein as a "light valve") or fiber optic light sources. By appropriately modulating liquid crystals, light may be selectively controlled so as to permit light to contact selected regions of the substrate. Alternatively, synthesis may take place on the end of a series of optical fibers to which light is selectively applied. Other means of controlling the location of light exposure will be apparent to those of skill in the art.

The substrate may be irradiated either in contact or not in contact with a solution (not shown) and is, preferably, irradiated in contact with a solution. The solution contains reagents to prevent the by-products formed by irradiation from interfering with synthesis of the polymer according to some embodiments. Such by-products might include, for example, carbon dioxide, nitrosocarbonyl compounds, styrene derivatives, indole derivatives, and products of their photochemical reactions. Alternatively, the solution may contain reagents used to match the index of refraction of the substrate. Reagents added to the solution may further include, for example, acidic or basic buffers, thiols, substituted hydrazines and hydroxylamines, reducing agents (e.g., NADH) or reagents known to react with a given functional group (e.g., aryl nitroso+glyoxylic acid→aryl formhydroxamate+CO_2).

Either concurrently with or after the irradiation step, the linker molecules are washed or otherwise contacted with a first monomer, illustrated by "A" in regions **12***a* and **12***b* in FIG. 2. The first monomer reacts with the activated functional groups of the linkage molecules which have been exposed to light. The first monomer, which is preferably an amino acid, is also provided with a photoprotective group. The photoprotective group on the monomer may be the same as or different than the protective group used in the linkage molecules, and may be selected from any of the above-described protective groups. In one embodiment, the protective groups for the A monomer is selected from the group NBOC and NVOC.

As shown in FIG. 3, the process of irradiating is thereafter repeated, with a mask repositioned so as to remove linkage protective groups and expose functional groups in regions **14***a* and **14***b* which are illustrated as being regions which were protected in the previous masking step. As an alternative to repositioning of the first mask, in many embodiments a second mask will be utilized. In other alternative embodiments, some steps may provide for illuminating a common region in successive steps. As shown in FIG. 3, it may be desirable to provide separation between irradiated regions. For example, separation of about 1–5 μm may be appropriate to account for alignment tolerances.

As shown in FIG. 4, the substrate is then exposed to a second protected monomer "B," producing B regions **16***a* and **16***b*. Thereafter, the substrate is again masked so as to remove the protective groups and expose reactive groups on A region **12***a* and B region **16***b*. The substrate is again exposed to monomer B, resulting in the formation of the structure shown in FIG. 6. The dimers B-A and B-B have been produced on the substrate.

A subsequent series of masking and contacting steps similar to those described above with A (not shown) provides the structure shown in FIG. 7. The process provides all possible dimers of B and A, i.e., B-A, A-B, A-A, and B-B.

The substrate, the area of synthesis, and the area for synthesis of each individual polymer could be of any size or shape. For example, squares, ellipsoids, rectangles, triangles, circles, or portions thereof, along with irregular geometric shapes, may be utilized. Duplicate synthesis areas may also be applied to a single substrate for purposes of redundancy.

In one embodiment the regions **12***a*, **12***b* and **16***a*, **16***b* on the substrate will have a surface area of between about 1 cm^2 and 10^{-10} cm^2. In some embodiments the regions **12***a*, **12***b* and **16***a*, **16***b* have areas of less than about 10^{-1} cm^2, 10^{-2} cm^2, 10^{-3} cm^2, 10^{-4} cm^2, 10^{-5} cm^2, 10^{-6} cm^2, 10^{-7} cm^2, 10^{-8} cm^2, or 10^{-10} cm^2. In a preferred embodiment, the regions **12***a*, **12***b* and **16***a*, **16***b* are between about 10×10 μm and 500×500 μm.

In some embodiments a single substrate supports more than about 10 different monomer sequences and perferably more than about 100 different monomer sequences, although in some embodiments more than about 10^3, 10^4, 10^5, 10^6, 10^7, or 10^8 different sequences are provided on a substrate. Of course, within a region of the substrate in which a monomer sequence is synthesized, it is preferred that the monomer sequence be substantially pure. In some embodiments, regions of the substrate contain polymer sequences which are at least about 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 97% 98% or 99% pure.

According to some embodiments, several sequences are intentionally provided within a single region so as to provide an initial screening for biological activity, after which materials within regions exhibiting significant binding are further evaluated.

IV. Details of One Embodiment of a Reactor System

FIG. 8A schematically illustrates a preferred embodiment of a reactor system **100** for synthesizing polymers on the prepared substrate in accordance with one aspect of the invention. The reactor system includes a body **102** with a cavity **104** on a surface thereof. In preferred embodiments the cavity **104** is between about 50 and 1000 μm deep with a depth of about 500 μm preferred.

The bottom of the cavity is preferably provided with an array of ridges **106** which extend both into the plane of the Figure and parallel to the plane of the Figure. The ridges are preferably about 50 to 200 μm deep and spaced at about 2 to 3 mm. The purpose of the ridges is to generate turbulent flow for better mixing. The bottom surface of the cavity is preferably light absorbing so as to prevent reflection of impinging light.

A substrate **112** is mounted above the cavity **104**. The substrate is provided along its bottom surface **114** with a photoremovable protective group such as NVOC with or without an intervening linker molecule. The substrate is preferably transparent to a wide spectrum of light, but in some embodiments is transparent only at a wavelength at which the protective group may be removed (such as UV in the case of NVOC). The substrate in some embodiments is a conventional microscope glass slide or cover slip. The substrate is preferably as thin as possible, while still providing adequate physical support. Preferably, the substrate is less than about 1 mm thick, more preferably less than 0.5 mm thick, more preferably less than 0.1 mm thick, and most preferably less than 0.05 mm thick. In alternative preferred embodiments, the substrate is quartz or silicon.

The substrate and the body serve to seal the cavity except for an inlet port **108** and an outlet port **110**. The body and the substrate may be mated for sealing in some embodiments with one or more gaskets. According to a preferred embodiment, the body is provided with two concentric gaskets and the intervening space is held at vacuum to ensure mating of the substrate to the gaskets.

Fluid is pumped through the inlet port into the cavity by way of a pump **116** which may be, for example, a model no. B-120-S made by Eldex Laboratories. Selected fluids are circulated into the cavity by the pump, through the cavity, and out the outlet for recirculation or disposal. The reactor may be subjected to ultrasonic radiation and/or heated to aid in agitation in some embodiments.

Above the substrate **112**, a lens **120** is provided which may be, for example, a 2"100 mm focal length fused silica lens. For the sake of a compact system, a reflective mirror **122** may be provided for directing light from a light source **124** onto the substrate. Light source **124** may be, for example, a Xe(Hg) light source manufactured by Oriel and having model no. 66024. A second lens **126** may be provided for the purpose of projecting a mask image onto the substrate in combination with lens 120. This form of lithography is referred to herein as projection printing. As will be apparent from this disclosure, proximity printing and the like may also be used according to some embodiments.

Light from the light source is permitted to reach only selected locations on the substrate as a result of mask **128**. Mask **128** may be, for example, a glass slide having etched chrome thereon. The mask 128 in one embodiment is provided with a grid of transparent locations and opaque locations. Such masks may be manufactured by, for example, Photo Sciences, Inc. Light passes freely through the transparent regions of the mask, but is reflected from or absorbed by other regions. Therefore, only selected regions of the substrate are exposed to light.

As discussed above, light valves (LCD's) may be used as an alternative to conventional masks to selectively expose regions of the substrate. Alternatively, fiber optic faceplates such as those available from Schott Glass, Inc, may be used for the purpose of contrast enhancement of the mask or as the sole means of restricting the region to which light is applied. Such faceplates would be placed directly above or on the substrate in the reactor shown in FIG. 8A. In still further embodiments, flys-eye lenses, tapered fiber optic faceplates, or the like, may be used for contrast enhancement.

In order to provide for illumination of regions smaller than a wavelength of light, more elaborate techniques may be utilized. For example, according to one preferred embodiment, light is directed at the substrate by way of molecular microcrystals on the tip of, for example, micropipettes. Such devices are disclosed in Lieberman et al., "A Light Source Smaller Than the Optical Wavelength," *Science* (1990) 247:59–61, which is incorporated herein by reference for all purposes.

In operation, the substrate is placed on the cavity and sealed thereto. All operations in the process of preparing the substrate are carried out in a room lit primarily or entirely by light of a wavelength outside of the light range at which the protective group is removed. For example, in the case of NVOC, the room should be lit with a conventional dark room light which provides little or no UV light. All operations are preferably conducted at about room temperature.

A first, deprotection fluid (without a monomer) is circulated through the cavity. The solution preferably is of 5 mM sulfuric acid in dioxane solution which serves to keep exposed amino groups protonated and decreases their reactivity with photolysis by-products. Absorptive materials such as N,N-diethylamino 2,4-dinitrobenzene, for example, may be included in the deprotection fluid which serves to absorb light and prevent reflection and unwanted photolysis.

The slide is, thereafter, positioned in a light raypath from the mask such that first locations on the substrate are illuminated and, therefore, deprotected. In preferred embodiments the substrate is illuminated for between about 1 and 15 minutes with a preferred illumination time of about 10 minutes at 10–20 mW/cm^2 with 365 nm light. The slides are neutralized (i.e., brought to a pH of about 7) after photolysis with, for example, a solution of di-isopropylethylamine (DIEA) in methylene chloride for about 5 minutes.

The first monomer is then placed at the first locations on the substrate. After irradiation, the slide is removed, treated in bulk, and then reinstalled in the flow cell. Alternatively, a fluid containing the first monomer, preferably also protected by a protective group, is circulated through the cavity by way of pump 116. If, for example, it is desired to attach the amino acid Y to the substrate at the first locations, the amino acid Y (bearing a protective group on its α-nitrogen), along with reagents used to render the monomer reactive, and/or a carrier, is circulated from a storage container 118, through the pump, through the cavity, and back to the inlet of the pump.

The monomer carrier solution is, in a preferred embodiment, formed by mixing of a first solution (referred to herein as solution "A") and a second solution (referred to herein as solution "B"). Table 2 provides an illustration of a mixture which may be used for solution A.

TABLE 2

Representative Monomer Carrier Solution "A"
100 mg NVOC amino protected amino acid
37 mg HOBT (1-Hydroxybenzotriazole)
250 μl DMF (Dimethylformamide)
86 μl DIEA (Diisopropylethylamine)

The composition of solution B is illustrated in Table 3. Solutions A and B are mixed and allowed to react at room temperature for about 8 minutes, then diluted with 2 ml of DMF, and 500 μl are applied to the surface of the slide or the solution is circulated through the reactor system and allowed to react for about 2 hours at room temperature. The slide is then washed with DMF, methylene chloride and ethanol.

TABLE 3

Representative Monomer Carrier Solution "B"
250 μl DMF
111 mg BOP (Benzotriazolyl-n-oxy-tris(dimethylamino) phosphoniumhexafluorophosphate)

As the solution containing the monomer to be attached is circulated through the cavity, the amino acid or other monomer will react at its carboxy terminus with amino groups on the regions of the substrate which have been deprotected. Of course, while the invention is illustrated by way of circulation of the monomer through the cavity, the invention could be practiced by way of removing the slide from the reactor and submersing it in an appropriate monomer solution.

After addition of the first monomer, the solution containing the first amino acid is then purged from the system. After circulation of a sufficient amount of the DMF/methylene chloride such that removal of the amino acid can be assured (e.g., about 50× times the volume of the cavity and carrier lines), the mask or substrate is repositioned, or a new mask is utilized such that second regions on the substrate will be exposed to light and the light 124 is engaged for a second exposure. This will deprotect second regions on the substrate and the process is repeated until the desired polymer sequences have been synthesized.

The entire derivatized substrate is then exposed to a receptor of interest, preferably labeled with, for example, a fluorescent marker, by circulation of a solution or suspension of the receptor through the cavity or by contacting the surface of the slide in bulk. The receptor will preferentially bind to certain regions of the substrate which contain complementary sequences.

Antibodies are typically suspended in what is commonly referred to as "supercocktail," which may be, for example, a solution of about 1% BSA (bovine serum albumin), 0.5% Tween™ non-ionic detergent in PBS (phosphate buffered saline) buffer. The antibodies are diluted into the supercocktail buffer to a final concentration of, for example, about 0.1 to 4 μg/ml.

FIG. 8B illustrates an alternative preferred embodiment of the reactor shown in FIG. 8A. According to this embodiment, the mask 128 is placed directly in contact with the substrate. Preferably, the etched portion of the mask is placed face down so as to reduce the effects of light dispersion. According to this embodiment, the imaging lenses 120 and 126 are not necessary because the mask is brought into close proximity with the substrate.

For purposes of increasing the signal-to-noise ratio of the technique, some embodiments of the invention provide for exposure of the substrate to a first labeled or unlabeled receptor followed by exposure of a labeled, second receptor (e.g., an antibody) which binds at multiple sites on the first receptor. If, for example, the first receptor is an antibody derived from a first species of an animal, the second receptor is an antibody derived from a second species directed to epitopes associated with the first species. In the case of a mouse antibody, for example, fluorescently labeled goat antibody or antiserum which is antimouse may be used to bind at multiple sites on the mouse antibody, providing several times the fluorescence compared to the attachment of a single mouse antibody at each binding site. This process may be repeated again with additional antibodies (e.g., goat-mouse-goat, etc.) for further signal amplification.

In preferred embodiments an ordered sequence of masks is utilized. In some embodiments it is possible to use as few as a single mask to synthesize all of the possible polymers of a given monomer set.

If, for example, it is desired to synthesize all 16 dinucleotides from four bases, a 1 cm square synthesis region is divided conceptually into 16 boxes, each 0.25 cm wide. Denote the four monomer units by A, B, C, and D. The first reactions are carried out in four vertical columns, each 0.25 cm wide. The first mask exposes the left-most column of boxes, where A is coupled. The second mask exposes the next column, where B is coupled; followed by a third mask, for the C column; and a final mask that exposes the right-most column, for D. The first, second, third, and fourth masks may be a single mask translated to different locations.

The process is repeated in the horizontal direction for the second unit of the dimer. This time, the masks allow exposure of horizontal rows, again 0.25 cm wide. A, B, C, and D are sequentially coupled using masks that expose horizontal fourths of the reaction area. The resulting substrate contains all 16 dinucleotides of four bases.

The eight masks used to synthesize the dinucleotide are related to one another by translation or rotation. In fact, one mask can be used in all eight steps if it is suitably rotated and translated. For example, in the example above, a mask with a single transparent region could be sequentially used to expose each of the vertical columns, translated 90°, and then sequentially used to allow exposure of the horizontal rows.

Tables 4 and 5 provide a simple computer program in Quick Basic for planning a masking program and a sample output, respectively, for the synthesis of a polymer chain of three monomers ("residues") having three different monomers in the first level, four different monomers in the second level, and five different monomers in the third level in a striped pattern. The output of the program is the number of cells, the number of "stripes" (light regions) on each mask, and the amount of translation required for each exposure of the mask.

TABLE 4

Mask Strategy Program

```
DEFINT A-Z
DIM b(20), w(20), 1(500)
FS = "LPT1:"
OPEN fS FOR OUTPUT AS #1
jmax = 3          'Number of residues
b(1) = 3: b(2) = 4: b(3) = 5     'Number of building blocks for res 1,2,3
g = 1: lmax(1) = 1
FOR j = 1 TO jmax: g= g * b(j): NEXT j
w(0) = 0: w(1) = g / b(1)
PRINT #1, "MASK2.BAS ", DATES, TIMES: PRINT #1,
PRINT #1, USING "Number of residues=##"; jmax
FOR j = 1 TO jmax
PRINT #1, USING "     Residue ##      ## building blocks"; j; b(j)
NEXT j
PRINT #1, "
PRINT #1, USING "Number of cells=####"; g: PRINT #1,
FOR j = 2 TO jmax
lmax(j) = lmax(j - 1) * b(j - 1)
w(j) = w(j - 1) / b(j)
NEXT j
FOR j = 1 TO jmax
PRINT #1, USING "Mask for residue ##"; j: PRINT #1,
PRINT #1, USING "   Number of stripes=###"; lmax(j)
PRINT #1, USING "   Width of each stripe=###"; w(j)
FOR 1 = 1 TO lmax(j)
a = 1 + (1 - 1) * w(j - 1)
ae = a + w(j) - 1
PRINT #1, USING "   Stripe ## begins at location ### and ends at ###"; l; a; ae
NEXT 1
PRINT #1,
PRINT #1, USING "   For each of ## building blocks, translate mask by ##
cell(s)"; b(j); w(j),
PRINT #1, : PRINT #1, : PRINT #1,
NEXT j
```

TABLE 5

Masking Strategy Output

```
Number of residues= 3
     Residue 1      3 building blocks
     Residue 2      4 building blocks
     Residue 3      5 building blocks
```

TABLE 5-continued

Masking Strategy Output

```
Number of cells= 60
Mask for residue 1
    Number of stripes= 1
    Width of each stripe= 20
    Stripe 1 begins at location 1 and ends at 20
    For each of 3 building blocks, translate mask by 20 cell(s)
Mask for residue 2
    Number of stripes= 3
    Width of each stripe= 5
    Stripe 1 begins at location 1 and ends at 5
    Stripe 2 begins at location 21 and ends at 25
    Stripe 3 begins at location 41 and ends at 45
    For each of 4 building blocks, translate mask by 5 cell(s)
Mask for residue 3
    Number of stripes= 12
    Width of each stripe= 1
    Stripe 1 begins at location 1 and ends at 1
    Stripe 2 begins at location 6 and ends at 6
    Stripe 3 begins at location 11 and ends at 11
    Stripe 4 begins at location 16 and ends at 16
    Stripe 5 begins at location 21 and ends at 21
    Stripe 6 begins at location 26 and ends at 26
    Stripe 7 begins at location 31 and ends at 31
    Stripe 8 begins at location 36 and ends at 36
    Stripe 9 begins at location 41 and ends at 41
    Stripe 10 begins at location 46 and ends at 46
    Stripe 11 begins at location 51 and ends at 51
    Stripe 12 begins at location 56 and ends at 56
    For each of 5 building blocks, translate mask by 1 cell(s)
```

V. Details of One Embodiment of A Fluorescent Detection Device

FIG. 9 illustrates a fluorescent detection device for detecting fluorescently labeled receptors on a substrate. A substrate 112 is placed on an x/y translation table 202. In a preferred embodiment the x/y translation table is a model no. PM500-A1 manufactured by Newport Corporation. The x/y translation table is connected to and controlled by an appropriately programmed digital computer 204 which may be, for example, an appropriately programmed IBM PC/AT or AT compatible computer. Of course, other computer systems, special purpose hardware, or the like could readily be substituted for the AT computer used herein for illustration. Computer software for the translation and data collection functions described herein can be provided based on commercially available software including, for example, "Lab Windows" licensed by National Instruments, which is incorporated herein by reference for all purposes.

The substrate and x/y translation table are placed under a microscope 206 which includes one or more objectives 208. Light (about 488 nm) from a laser 210, which in some embodiments is a model no. 2020-05 argon ion laser manufactured by Spectraphysics, is directed at the substrate by a dichroic mirror 207 which passes greater than about 520 nm light but reflects 488 nm light. Dichroic mirror 207 may be, for example, a model no. FT510 manufactured by Carl Zeiss. Light reflected from the mirror then enters the microscope 206 which may be, for example, a model no. Axioscop 20 manufactured by Carl Zeiss. Fluorescein-marked materials on the substrate will fluoresce >488 nm light, and the fluoresced light will be collected by the microscope and passed through the mirror. The fluorescent light from the substrate is then directed through a wavelength filter 209 and, thereafter through an aperture plate 211. Wavelength filter 209 may be, for example, a model no. OG530 manufactured by Melles Griot and aperture plate 211 may be, for example, a model no. 477352/477380 manufactured by Carl Zeiss.

The fluoresced light then enters a photomultiplier tube 212 which in some embodiments is a model no. R943-02 manufactured by Hamamatsu, the signal is amplified in preamplifier 214 and photons are counted by photon counter 216. The number of photons is recorded as a function of the location in the computer 204. Pre-Amp 214 may be, for example, a model no. SR440 manufactured by Stanford Research Systems and photon counter 216 may be a model no. SR400 manufactured by Stanford Research Systems. The substrate is then moved to a subsequent location and the process is repeated. In preferred embodiments the data are acquired every 1 to 100 μm with a data collection diameter of about 0.8 to 10 μm preferred. In embodiments with sufficiently high fluorescence, a CCD (change coupled device) detector with broadfield illumination is utilized.

By counting the number of photons generated in a given area in response to the laser, it is possible to determine where fluorescent marked molecules are located on the substrate. Consequently, for a slide which has a matrix of polypeptides, for example, synthesized on the surface thereof, it is possible to determine which of the polypeptides is complementary to a fluorescently marked receptor.

According to preferred embodiments, the intensity and duration of the light applied to the substrate is controlled by varying the laser power and scan stage rate for improved signal-to-noise ratio by maximizing fluorescence emission and minimizing background noise.

While the detection apparatus has been illustrated primarily herein with regard to the detection of marked receptors, the invention will find application in other areas. For example, the detection apparatus disclosed herein could be used in the fields of catalysis, DNA or protein gel scanning, and the like.

VI. Determination of Relative Binding Strength of Receptors

The signal-to-noise ratio of the present invention is sufficiently high that not only can the presence or absence of a receptor on a ligand be detected, but also the relative binding affinity of receptors to a variety of sequences can be determined.

In practice it is found that a receptor will bind to several peptide sequences in an array, but will bind much more strongly to some sequences than others. Strong binding affinity will be evidenced herein by a strong fluorescent or radiographic signal since many receptor molecules will bind in a region of a strongly bound ligand. Conversely, a weak binding affinity will be evidenced by a weak fluorescent or radiographic signal due to the relatively small number of receptor molecules which bind in a particular region of a substrate having a ligand with a weak binding affinity for the receptor. Consequently, it becomes possible to determine relative binding avidity (or affinity in the case of univalent interactions) of a ligand herein by way of the intensity of a fluorescent or radiographic signal in a region containing that ligand.

Semiquantitative data on affinities might also be obtained by varying washing conditions and concentrations of the receptor. This would be done by comparison to known ligand receptor pairs, for example.

VII. Examples

The following examples are provided to illustrate the efficacy of the inventions herein. All operations were conducted at about ambient temperatures and pressures unless indicated to the contrary.

A. Slide Preparation

Before attachment of reactive groups it is preferred to clean the substrate which is, in a preferred embodiment a glass substrate such as a microscope slide or cover slip. According to one embodiment the slide is soaked in an alkaline bath consisting of, for example, 1 liter of 95% ethanol with 120 ml of water and 120 grams of sodium hydroxide for 12 hours. The slides are then washed under running water and allowed to air dry, and rinsed once with a solution of 95% ethanol.

The slides are then aminated with, for example, aminopropyltriethoxysilane for the purpose of attaching amino groups to the glass surface on linker molecules, although any omega functionalized silane could also be used for this purpose. In one embodiment 0.1% aminopropyltriethoxysilane is utilized, although solutions with concentrations from $10^{-7}\%$ to 10% may be used, with about $10^{-3}\%$ to 2% preferred. A 0.1% mixture is prepared by adding to 100 ml of a 95% ethanol/5% water mixture, 100 microliters (μl) of aminopropyltriethoxysilane. The mixture is agitated at about ambient temperature on a rotary shaker for about 5 minutes. 500 μl of this mixture is then applied to the surface of one side of each cleaned slide. After 4 minutes, the slides are decanted of this solution and rinsed three times by dipping in, for example, 100% ethanol.

After the plates dry, they are placed in a 110°-120° C. vacuum oven for about 20 minutes, and then allowed to cure at room temperature for about 12 hours in an argon environment. The slides are then dipped into DMF (dimethylformamide) solution, followed by a thorough washing with methylene chloride.

The aminated surface of the slide is then exposed to about 500 μl of, for example, a 30 millimolar (mM) solution of NVOC-GABA (gamma amino butyric acid) NHS (N-hydroxysuccinimide) in DMF for attachment of a NVOC-GABA to each of the amino groups.

The surface is washed with, for example, DMF, methylene chloride, and ethanol.

Any unreacted aminopropyl silane on the surface—that is, those amino groups which have not had the NVOC-GABA attached—are now capped with acetyl groups (to prevent further reaction) by exposure to a 1:3 mixture of acetic anhydride in pyridine for 1 hour. Other materials which may perform this residual capping function include trifluoroacetic anhydride, formicacetic anhydride, or other reactive acylating agents. Finally, the slides are washed again with DMF, methylene chloride, and ethanol.

B. Synthesis of Eight Trimers of "A" and "B"

FIG. 10 illustrates a possible synthesis of the eight trimers of the two-monomer set: gly, phe (represented by "A" and "B," respectively). A glass slide bearing silane groups terminating in 6-nitroveratryloxycarboxamide (NVOC-NH) residues is prepared as a substrate. Active esters (pentafluorophenyl, OBt, etc.) of gly and phe protected at the amino group with NVOC are prepared as reagents. While not pertinent to this example, if side chain protecting groups are required for the monomer set, these must not be photoreactive at the wavelength of light used to protect the primary chain.

For a monomer set of size n, n×l cycles are required to synthesize all possible sequences of length l. A cycle consists of:

1. Irradiation through an appropriate mask to expose the amino groups at the sites where the next residue is to be added, with appropriate washes to remove the by-products of the deprotection.
2. Addition of a single activated and protected (with the same photochemically-removable group) monomer, which will react only at the sites addressed in step 1, with appropriate washes to remove the excess reagent from the surface.

The above cycle is repeated for each member of the monomer set until each location on the surface has been extended by one residue in one embodiment. In other embodiments, several residues are sequentially added at one location before moving on to the next location. Cycle times will generally be limited by the coupling reaction rate, now as short as 20 min in automated peptide synthesizers. This step is optionally followed by addition of a protecting group to stabilize the array for later testing. For some types of polymers (e.g., peptides), a final deprotection of the entire surface (removal of photoprotective side chain groups) may be required.

More particularly, as shown in FIG. 10A, the glass 20 is provided with regions 22, 24, 26, 28, 30, 32, 34, and 36. Regions 30, 32, 34, and 36 are masked, as shown in FIG. 10B and the glass is irradiated and exposed to a reagent containg "A" (e.g., gly), with the resulting structure shown in FIG. 10C. Thereafter, regions 22, 24, 26, and 28 are masked, the glass is irradiated (as shown in FIG. 10D) and exposed to a reagent containing "B" (e.g., phe), with the resulting structure shown in FIG. 10E. The process proceeds, consecutively masking and exposing the sections as shown until the structure shown in FIG. 10M is obtained. The glass is irradiated and the terminal groups are, optionally, capped by acetylation. As shown, all possible trimers of gly/phe are obtained.

In this example, no side chain protective group removal is necessary. If it is desired, side chain deprotection may be accomplished by treatment with ethanedithiol and trifluoroacetic acid.

In general, the number of steps needed to obtain a particular polymer chain is defined by:

$$n \times l \qquad (1)$$

where:

n=the number of monomers in the basis set of monomers, and

l=the number of monomer units in a polymer chain.

Conversely, the synthesized number of sequences of length l will be:

$$n^l. \qquad (2)$$

Of course, greater diversity is obtained by using masking strategies which will also include the synthesis of polymers having a length of less than l. If, in the extreme case, all polymers having a length less than or equal to l are synthesized, the number of polymers synthesized will be:

$$n^l+n^{l-1}+ \ldots +n^1. \qquad (3)$$

The maximum number of lithographic steps needed will generally be n for each "layer" of monomers, i.e., the total number of masks (and, therefore, the number of lithographic steps) needed will be n×l. The size of the transparent mask regions will vary in accordance with the area of the substrate available for synthesis and the number of sequences to be formed. In general, the size of the synthesis areas will be:

size of synthesis areas=(A)/(Sequences)

where:

A is the total area available for synthesis; and

Sequences is the number of sequences desired in the area.

It will be appreciated by those of skill in the art that the above method could readily be used to simultaneously produce thousands or millions of oligomers on a substrate using the photolithographic techniques disclosed herein. Consequently, the method results in the ability to practically test large numbers of, for example, di, tri, tetra, penta, hexa, hepta, octapeptides, dodecapeptides, or larger polypeptides (or correspondingly, polynucleotides).

The above example has illustrated the method by way of a manual example. It will of course be appreciated that automated or semi-automated methods could be used. The substrate would be mounted in a flow cell for automated addition and removal of reagents, to minimize the volume of reagents needed, and to more carefully control reaction conditions. Successive masks could be applied manually or automatically.

Synthesis of a Dimer of an Aminopropyl Group and a Fluorescent Group

In synthesizing the dimer of an aminopropyl group and a fluorescent group, a functionalized durapore membrane was used as a substrate. The durapore membrane was a polyvinylidine difluoride with aminopropyl groups. The aminopropyl groups were protected with the DDZ group by reaction of the carbonyl chloride with the amino groups, a reaction readily known to those of skill in the art. The surface bearing these groups was placed in a solution of THF and contacted with a mask bearing a checkerboard pattern of 1 mm opaque and transparent regions. The mask was exposed to ultraviolet light having a wavelength down to at least about 280 nm for about 5 minutes at ambient temperature, although a wide range of exposure times and temperatures may be appropriate in various embodiments of the invention. For example, in one embodiment, an exposure time of between about 1 and 5000 seconds may be used at process temperatures of between −70° and +50° C.

In one preferred embodiment, exposure times of between about 1 and 500 seconds at about ambient pressure are used. In some preferred embodiments, pressure above ambient is used to prevent evaporation.

The surface of the membrane was then washed for about 1 hour with a fluorescent label which included an active ester bound to a chelate of a lanthanide. Wash times will vary over a wide range of values from about a few minutes to a few hours. These materials fluoresce in the red and the green visible region. After the reaction with the active ester in the fluorophore was complete, the locations in which the fluorophore was bound could be visualized by exposing them to ultraviolet light and observing the red and the green fluorescence. It was observed that the derivatized regions of the substrate closely corresponded to the original pattern of the mask.

D. Demonstration of Signal Capability

Signal detection capability was demonstrated using a low-level standard fluorescent bead kit manufactured by Flow Cytometry Standards and having model no. 824. This kit includes 5.8 μm diameter beads, each impregnated with a known number of fluorescein molecules.

One of the beads was placed in the illumination field on the scan stage as shown in FIG. 9 in a field of a laser spot which was initially shuttered. After being positioned in the illumination field, the photon detection equipment was turned on. The laser beam was unblocked and it interacted with the particle bead, which then fluoresced. Fluorescence curves of beads impregnated with 7,000 and 13,000; fluorescein molecules, are shown in FIGS. 11A and 11B respectively. On each curve, traces for beads without fluorescein molecules are also shown. These experiments were performed with 488 nm excitation, with 100 μW of laser power. The light was focused through a 40 power 0.75 NA objective.

The fluorescence intensity in all cases started off at a high value and then decreased exponentially. The falloff in intensity is due to photobleaching of the fluorescein molecules. The traces of beads without fluorescein molecules are used for background subtraction. The difference in the initial exponential decay between labeled and nonlabeled beads is integrated to give the total number of photon counts, and this number is related to the number of molecules per bead. Therefore, it is possible to deduce the number of photons per fluorescein molecule that can be detected. For the curves illustrated in FIG. 11A and 11B, this calculation indicates the radiation of about 40 to 50 photons per fluorescein molecule are detected.

E. Determination of the Number of Molecules Per Unit Area

Aminopropylated glass microscope slides prepared according to the methods discussed above were utilized in order to establish the density of labeling of the slides. The free amino termini of the slides were reacted with FITC (fluorescein isothiocyanate) which forms a covalent linkage with the amino group. The slide is then scanned to count the number of fluorescent photons generated in a region which, using the estimated 40–50 photons per fluorescent molecule, enables the calculation of the number of molecules which are on the surface per unit area.

A slide with aminopropyl silane on its surface was immersed in a 1 mM solution of FITC in DMF for 1 hour at about ambient temperature. After reaction, the slide was washed twice with DMF and then washed with ethanol, water, and then ethanol again. It was then dried and stored in the dark until it was ready to be examined.

Through the use of curves similar to those shown in FIG. 11A and 11B, and by integrating the fluorescent counts under the exponentially decaying signal, the number of free amino groups on the surface after derivatization was determined. It was determined that slides with labeling densities of 1 fluorescein per $10^3\times10^3$ to $\sim2\times2$ nm could be reproducibly made as the concentration of aminopropyltriethoxysilane varied from 10^{-5}% to 10^{-1}%.

F. Removal of NVOC and Attachment of A Fluorescent Marker

NVOC-GABA groups were attached as described above. The entire surface of one slide was exposed to light so as to expose a free amino group at the end of the gamma amino butyric acid. This slide, and a duplicate which was not exposed, were then exposed to fluorescein isothiocyanate (FITC).

FIG. 12A illustrates the slide which was not exposed to light, but which was exposed to FITC. The units of the x axis are time and the units of the y axis are counts. The trace contains a certain amount of background fluorescence. The duplicate slide was exposed to 350 nm broadband illumination for about 1 minute (12 mW/cm^2, ~350 nm illumination), washed and reacted with FITC. The fluorescence curves for this slide are shown in FIG. 12B. A large increase in the level of fluorescence is observed, which indicates photolysis has exposed a number of amino groups on the surface of the slides for attachment of a fluorescent marker.

G. Use of a Mask in Removal of NVOC

The next experiment was performed with a 0.1% aminopropylated slide. Light from a Hg—Xe arc lamp was imaged onto the substrate through a laser-ablated chrome-on-glass mask in direct contact with the substrate.

This slide was illuminated for approximately 5 minutes, with 12 mW of 350 nm broadband light and then reacted with the 1 mM FITC solution. It was put on the laser detection scanning stage and a graph was plotted as a two-dimensional representation of position color-coded for fluorescence intensity. The fluorescence intensity (in counts) as a function of location is given on the color scale to the right of FIG. 13A for a mask having 100×100 μm squares.

The experiment was repeated a number of times through various masks. The fluorescence pattern for a 50 μm mask is illustrated in FIG. 13B, for a 20 μm mask in FIG. 13C, and for a 10 μm mask in FIG. 13D. The mask pattern is distinct down to at least about 10 μm squares using this lithographic technique.

H. Attachment of YGGFL and Subsequent Exposure to

Herz Antibody and Goat Antimouse

In order to establish that receptors to a particular polypeptide sequence would bind to a surface-bound peptide and be detected, Leu enkephalin was coupled to the surface and recognized by an antibody. A slide was derivatized with 0.1% amino propyl-triethoxysilane and protected with NVOC. A 500 μm checkerboard mask was used to expose the slide in a flow cell using backside contact printing. The Leu enkephalin sequence (H_2N-tyrosine,glycine,glycine,phenylalanine,leucine-CO_2H, otherwise referred to herein as YGGFL) was attached via its carboxy end to the exposed amino groups on the surface of the slide. The peptide was added in DMF solution with the BOP/HOBT/DIEA coupling reagents and recirculated through the flow cell for 2 hours at room temperature.

A first antibody, known as the Herz antibody, was applied to the surface of the slide for 45 minutes at 2 μg/ml in a supercocktail (containing 1% BSA and 1% ovalbumin also in this case). A second antibody, goat anti-mouse fluorescein conjugate, was then added at 2 μg/ml in the supercocktail buffer, and allowed to incubate for 2 hours. An image taken at 10 μm steps indicated that not only can deprotection be carried out in a well defined pattern, but also that (1) the method provides for successful coupling of peptides to the surface of the substrate, (2) the surface of a bound peptide is available for binding with an antibody, and (3) that the detection apparatus capabilities are sufficient to detect binding of a receptor.

I. Monomer-by-Monomer Formation of YGGFL and Subsequent Exposure to Labeled Antibody

Monomer-by-monomer synthesis of YGGFL and GGFL in alternate squares was performed on a slide in a checkerboard pattern and the resulting slide was exposed to the Herz antibody. This experiment and the results thereof are illustrated in FIGS. 14A, 14B, 15A, and 15B.

In FIG. 14A, a slide is shown which is derivatized with the aminopropyl group, protected in this case with t-BOC (t-butoxycarbonyl). The slide was treated with TFA to remove the t-BOC protecting group. E-aminocaproic acid, which was t-BOC protected at its amino group, was then coupled onto the aminopropyl groups. The aminocaproic acid serves as a spacer between the aminopropyl group and the peptide to be synthesized. The amino end of the spacer was deprotected and coupled to NVOC-leucine. The entire slide was then illuminated with 12 mW of 325 nm broadband illumination. The slide was then coupled with NVOC-phenylalanine and washed. The entire slide was again illuminated, then coupled to NVOC-glycine and washed. The slide was again illuminated and coupled to NVOC-glycine to form the sequence shown in the last portion of FIG. 14A.

As shown in FIG. 14B, alternating regions of the slide were then illuminated using a projection print using a 500×500 μm checkerboard mask; thus, the amino group of glycine was exposed only in the lighted areas. When the next coupling chemistry step was carried out, NVOC-tyrosine was added, and it coupled only at those spots which had received illumination. The entire slide was then illuminated to remove all the NVOC groups, leaving a checkerboard of YGGFL in the lighted areas and in the other areas, GGFL. The Herz antibody (which recognizes the YGGFL, but not GGFL) was then added, followed by goat anti-mouse fluorescein conjugate.

The resulting fluorescence scan is shown in FIG. 15A, and the color coding for the fluorescence intensity is again given on the right. Dark areas contain the tetrapeptide GGFL, which is not recognized by the Herz antibody (and thus there is no binding of the goat anti-mouse antibody with fluorescein conjugate), and in the red areas YGGFL is present. The YGGFL pentapeptide is recognized by the Herz antibody and, therefore, there is antibody in the lighted regions for the fluorescein-conjugated goat anti-mouse to recognize.

Similar patterns are shown for a 50 μm mask used in direct contact ("proximity print") with the substrate in FIG. 15B. Note that the pattern is more distinct and the corners of the checkerboard pattern are touching when the mask is placed in direct contact with the substrate (which reflects the increase in resolution using this technique).

J. Monomer-by-Monomer Synthesis of YGGFL and PGGFL

A synthesis using a 50 μm checkerboard mask similar to that shown in FIG. 15B was conducted. However, P was added to the GGFL sites on the substrate through an additional coupling step. P was added by exposing protected GGFL to light and subsequent exposure to P in the manner set forth above. Therefore, half of the regions on the substrate contained YGGFL and the remaining half contained PGGFL.

The fluorescence plot for this experiment is provided in FIG. 16. As shown, the regions are again readily discernable. This experiment demonstrates that antibodies are able to recognize a specific sequence and that the recognition is not length-dependent.

K. Monomer-by-Monomer Synthesis of YGGFL and YPGGFL

In order to further demonstrate the operability of the invention, a 50 μm checkerboard pattern of alternating YGGFL and YPGGFL was synthesized on a substrate using techniques like those set forth above. The resulting fluorescence plot is provided in FIG. 17. Again, it is seen that the antibody is clearly able to recognize the YGGFL sequence and does not bind significantly at the YPGGFL regions.

L. Synthesis of an Array of Sixteen Different Amino Acid Sequences and Estimation of Relative Binding Affinity to Herz Antibody

Using techniques similar to those set forth above, an array of 16 different amino acid sequences (replicated four times) was synthesized on each of two glass substrates. The sequences were synthesized by attaching the sequence NVOC-GFL across the entire surface of the slides. Using a series of masks, two layers of amino acids were then selectively applied to the substrate. Each region had dimensions of 0.25 cm×0.0625 cm. The first slide contained amino acid sequences containing only L amino acids while the second slide contained selected D amino acids. FIGS. 18A and 18B illustrate a map of the various regions on the first and second slides, respectively. The patterns shown in FIGS. 18A and 18B were duplicated four times on each slide. The slides were then exposed to the Herz antibody and fluorescein-labeled goat anti-mouse.

FIG. 19 is a fluorescence plot of the first slide, which contained only L amino acids. Red indicates strong binding (149,000 counts or more) while black indicates little or no binding of the Herz antibody (20,000 counts or less). The bottom right-hand portion of the slide appears "cut off" because the slide was broken during processing. The sequence YGGFL is clearly most strongly recognized. The sequences YAGFL and YSGFL also exhibit strong recognition of the antibody. By contrast, most of the remaining sequences show little or no binding. The four duplicate portions of the slide are extremely consistent in the amount of binding shown therein.

FIG. 20 is a fluorescence plot of the second slide. Again, strongest binding is exhibited by the YGGFL sequence. Significant binding is also detected to YaGFL, YsGFL, and YpGFL (where L-amino acids are identified by one upper case letter abbreviation, and D-amino acids are identified by one lower case letter abbreviation). The remaining sequences show less binding with the antibody. Note the low binding efficiency of the sequence yGGFL.

Table 6 lists the various sequences tested in order of relative fluorescence, which provides information regarding relative binding affinity.

TABLE 6

Apparent Binding to Herz Ab	
L-a.a. Set	D-a.a. Set
YGGFL	YGGFL
YAGFL	YaGFL
YSGFL	YsGFL
LGGFL	YpGFL
FGGFL	fGGFL
YPGFL	yGGFL
LAGFL	faGFL
FAGFL	wGGFL
WGGFL	yaGFL
	fpGFL
	waGFL

VIII. Illustrative Alternative Embodiment

According to an alternative embodiment of the invention, the methods provide for attaching to the surface a caged binding member which in its caged form has a relatively low affinity for other potentially binding species, such as receptors and specific binding substances.

According to this alternative embodiment, the invention provides methods for forming predefined regions on a surface of a solid support, wherein the predefined regions are capable of immobilizing receptors. The methods make use of caged binding members attached to the surface to enable selective activation of the predefined regions. The caged binding members are liberated to act as binding members ultimately capable of binding receptors upon selective activation of the predefined regions. The activated binding members are then used to immobilize specific molecules such as receptors on the predefined region of the surface. The above procedure is repeated at the same or different sites on the surface so as to provide a surface prepared with a plurality of regions on the surface containing, for example, the same or different receptors. When receptors immobilized in this way have a differential affinity for one or more ligands, screenings and assays for the ligands can be conducted in the regions of the surface containing the receptors.

The alternative embodiment may make use of novel caged binding members attached to the substrate. Caged (unactivated) members have a relatively low affinity for receptors of substances that specifically bind to uncaged binding members when compared with the corresponding affinities of activated binding members. Thus, the binding members are protected from reaction until a suitable source of energy is applied to the regions of the surface desired to be activated. Upon application of a suitable energy source, the caging groups labilize, thereby presenting the activated binding member. A typical energy source will be light.

Once the binding members on the surface are activated they may be attached to a receptor. The receptor chosen may be a monoclonal antibody, a nucleic acid sequence, a drug receptor, etc. The receptor will usually, though not always, be prepared so as to permit attaching it, directly or indirectly, to a binding member. For example, a specific binding substance having a strong binding affinity for the binding member and a strong affinity for the receptor or a conjugate of the receptor may be used to act as a bridge between binding members and receptors if desired. The method uses a receptor prepared such that the receptor retains its activity toward a particular ligand.

Preferably, the caged binding member attached to the solid substrate will be a photoactivatable biotin complex, i.e., a biotin molecule that has been chemically modified with photoactivatable protecting groups so that it has a significantly reduced binding affinity for avidin or avidin analogs than does natural biotin. In a preferred embodiment, the protecting groups localized in a predefined region of the surface will be removed upon application of a suitable source of radiation to give binding members, that are biotin or a functionally analogous compound having substantially the same binding affinity for avidin or avidin analogs as does biotin.

In another preferred embodiment, avidin or an avidin analog is incubated with activated binding members on the surface until the avidin binds strongly to the binding members. The avidin so immobilized on predefined regions of the surface can then be incubated with a desired receptor or conjugate of a desired receptor. The receptor will preferably be biotinylated, e.g., a biotinylated antibody, when avidin is immobilized on the predefined regions of the surface. Alternatively, a preferred embodiment will present an avidin/biotinylated receptor complex, which has been previously prepared, to activated binding members on the surface.

IX. Conclusion

The present inventions provide greatly improved methods and apparatus for synthesis of polymers on substrates. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments will be apparent to those of skill in the art upon reviewing the above description. By way of example, the invention has been described primarily with reference to the use of photoremovable protective groups, but it will be readily recognized by those of skill in the art that sources of radiation other than light could also be used. For example, in some embodiments it may be desirable to use protective groups which are sensitive to electron beam irradiation, x-ray irradiation, in combination with electron beam lithograph, or x-ray lithography techniques. Alternatively, the group could be removed by exposure to an electric current. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

What is claimed is:

1. A substrate with a surface comprising 10^3 or more groups of oligonucleotides with different, known sequences covalently attached to the surface in discrete known regions, said 10^3 or more groups of oligonucleotides occupying a total area of less than 1 cm^2 on said substrate, said groups of oligonucleotides having different nucleotide sequences.

2. The substrate as recited in claim 1 wherein said substrate comprises 10^4 or more different groups of oligonucleotide with known sequences covalently coupled to discrete known regions of said substrate.

3. The substrate as recited in claim 1 wherein said substrate comprises 10^5 or more different groups of oligonucleotides with known sequences in discrete known regions.

4. The substrate as recited in claim 1 wherein said substrate comprises 10^6 or more different groups of oligonucleotides with known sequences in discrete known regions.

5. The substrate as recited in claim 1 wherein said groups of oligonucleotides are at least 50% pure within said discrete known regions.

6. The substrate as recited in claim 1 wherein the groups of oligonucleotides are attached to the surface by a linker.

7. An array of more than 1,000 different groups of oligonucleotide molecules with known sequences covalently coupled to a surface of a substrate, said groups of oligonucleotide molecules each in discrete known regions and differing from other groups of oligonucleotide molecules in monomer sequence, each of said discrete known regions being an area of less than about 0.01 cm^2 and each discrete known region comprising oligonucleotides of known sequence, said different groups occupying a total area of less than 1 cm^2.

8. The array as recited in claim 7 wherein said area is less than 10,000 microns^2.

9. The array as recited in claim 7 made by the process of:

exposing a first region of said substrate to light to remove photoremovable groups from nucleic acids in said first region, and not exposing a second region of said surface to light;

covalently coupling a first nucleotide to said nucleic acids on said part of said substrate exposed to light, said first nucleotide covalently coupled to said photoremovable group;

exposing a part of said first region of said substrate to light, and not exposing another part of said first region of said substrate to light to remove said photoremovable groups;

covalently coupling a second nucleotide to said part of said first region exposed to light; and

repeating said steps of exposing said substrate to light and covalently coupling nucleotides until said more than 500 different groups of nucleotides are formed on said surface.

10. The array as recited in claim 7 comprising more than 10,000 groups of oligonucleotides of known sequences.

* * * * *

X

US005556752A

United States Patent [19]

Lockhart et al.

[11] **Patent Number: 5,556,752**

[45] **Date of Patent: Sep. 17, 1996**

[54] **SURFACE-BOUND, UNIMOLECULAR, DOUBLE-STRANDED DNA**

[75] Inventors: **David J. Lockhart**, Santa Clara, Calif.; **Dirk Vetter**, Freiburg, Germany; **Martin Diggelmann**, Niederdorf, Switzerland

[73] Assignee: **Affymetrix, Inc.**, Santa Clara, Calif.

[21] Appl. No.: **327,687**

[22] Filed: **Oct. 24, 1994**

[51] **Int. Cl.**^{6} **C12Q 1/68; C07H 21/00**

[52] **U.S. Cl.** 435/6; 536/23.1

[58] **Field of Search** 435/6; 536/23.1; 530/413

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,376,110	3/1983	David et al.	435/5
4,562,157	12/1985	Lowe et al.	435/287.2
4,728,502	3/1988	Hamill	422/116
5,143,854	9/1992	Pirrung et al.	436/518
5,288,514	2/1994	Ellman	165/155

FOREIGN PATENT DOCUMENTS

WO89/10977	11/1989	WIPO .
WO89/11548	11/1989	WIPO .
WO90/00626	1/1990	WIPO .
WO90/15070	12/1990	WIPO .
WO92/00091	1/1992	WIPO .

OTHER PUBLICATIONS

Duncan, C. H. etal (1988) Analytical Biochemistry 169: 104–108. "Affinity Chromatography of a Sequence-specific DNA binding protein using Teflon linked . . . ".

Ma, M. Y.-X. et al (1993) Biochemistry 32: 1751–1758. "Design & Synthesis of RNA Miniduplicates via a synthetic linker approach."Markiewicz, W T et al (1989) Nucleic Acids Research 17: 7149–7157. "Universal solid supports for the synthesis of oligonucleotides with 3'- PO_4s".

Ohlmeyer, M H J et al (1993) Proc. Natl. Acad. Sci. USA 90: 10922–10926 "Complex Synthetic Chemical Libraries Indexed with molecular Tags."Geysen, et al., *J. Immun. Meth.* 102:259–274 (1987).

Frank and Doring, *Tetrahedron*, 44:6031–6040 (1988).

Fodor et al., *Science*, 251:767–777 (1991).

Lam et al., *Nature*, 354:82–84 (1991).

Houghten et al., *Nature*, 354:84–86 (1991).

Galas et al., *Nucleic Acid Res.* 5(9):3157–3170 (1978).

Murphy et al., *Science* 262:1025–1029 (1993).

Lysov et al., *Dokl. Akad. Nauk SSSR*, 303:1508–1511 (1988) (See footnote provided, P. 436).

Bains et al., *J. Theor. Biol.*, 135:303–307 (1988).

Drmanac et al., Genomics, 4:114–128 (1989).

Strezoska et al., *Proc. Natl. Acad. Sci. USA*, 88:10089–10093 (1991).

Drmanac et al., *Science*, 260:1649–1652 (1993).

Needels, et al., *Proc. Natl. Acad. Sci., USA*, 90:10700–10704 (1993).

Scaria, P. V., et al. *J. of Biol. Chem.*, 266(9) : 5417–5423 (1993).

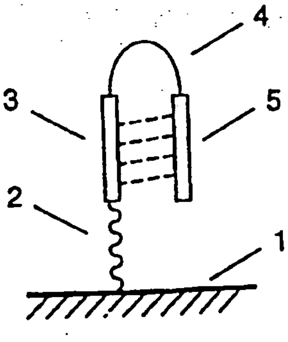
(List continued on next page.)

Primary Examiner—Mindy Fleisher
Assistant Examiner—Scott David Priebe
Attorney, Agent, or Firm—Townsend and Townsend and Crew LLP

[57] **ABSTRACT**

Libraries of unimolecular, double-stranded oligonucleotides on a solid support. These libraries are useful in pharmaceutical discovery for the screening of numerous biological samples for specific interactions between the double-stranded oligonucleotides, and peptides, proteins, drugs and RNA. In a related aspect, the present invention provides libraries of conformationally restricted probes on a solid support. The probes are restricted in their movement and flexibility using double-stranded oligonucleotides as scaffolding. The probes are also useful in various screening procedures associated with drug discovery and diagnosis. The present invention further provides methods for the preparation and screening of the above libraries.

6 Claims, 1 Drawing Sheet



OTHER PUBLICATIONS

Durand, M., et al., *Nucleic Acid Res.*, 18(21) : 6353–5469 (1990).

Famulok, M., et al., *Angew. Chem. Int. Ed. Engl.*, 31:979–988 (1992).

Chattopadhyaya, R., et al., *Nature*, 334:175–179 (1988).

Bock, L. C., et al., *Nature*, 355:564–566 (1992).

Parham, Peter, *Nature*, 360:300–301 (1992).

Tuerk, C., et al., *Science*, 249:505–510 (1990).

Mergny, J.–L., et al., *Nucleic Acids Res.*, 19(7) : 1521–1526 (1991).

Brossalina, E., et al., *J. Am. Chem. Soc.*, 115:796–797 (1993).

Härd, T., et al., *Biochemistry*, 29:959–965 (1990).

Cook, J., et al., *Analytical Biochemistry*, 190:331–339 (1990).

Cuniberti, C. et al., *Biophysical Chemistry*, 38:11–22 (1990).

Berman, H. M., et al., *Ann. Rev. Biophys. Bioeng.*, 10:87–114 (1981).

White et al. "Principles of Biochemistry" New York: McGraw–Hill, 1978 pp. 124–128.

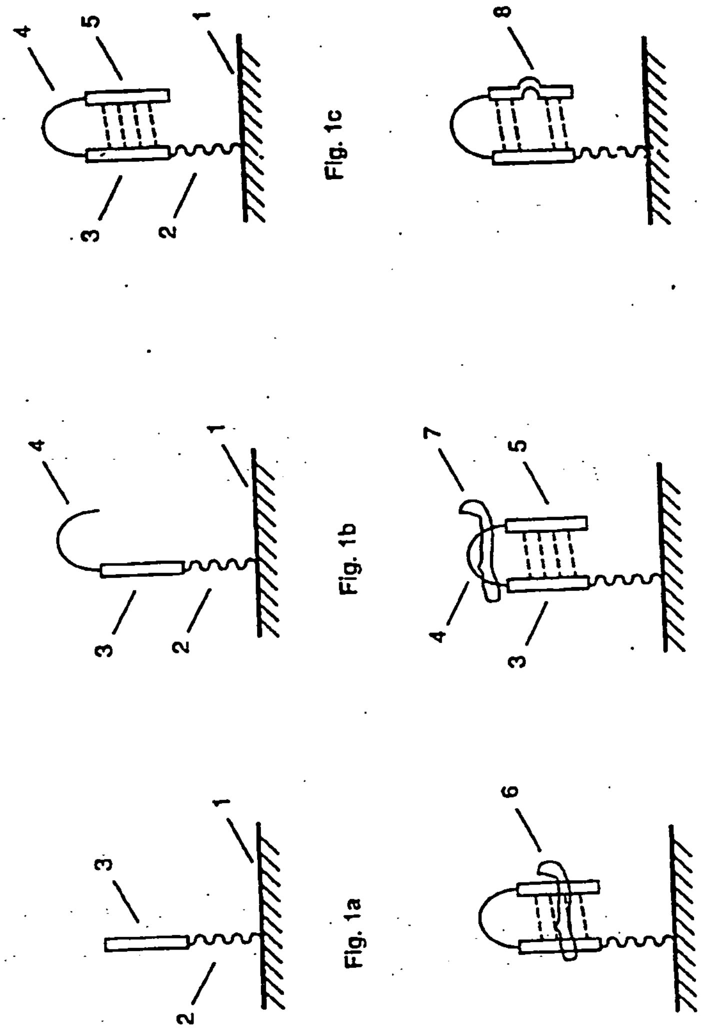


Fig. 1a

Fig. 1b

Fig. 1c

Fig. 1d

Fig. 1e

Fig. 1f

Figure 1a-1f

SURFACE-BOUND, UNIMOLECULAR, DOUBLE-STRANDED DNA

GOVERNMENT RIGHTS

Research leading to the invention was funded in part by NIH Grant No. R01HG00813-03 and the government may have certain fights to the invention.

BACKGROUND OF THE INVENTION

The present invention relates to the field of polymer synthesis and the use of polymer libraries for biological screening. More specifically, in one embodiment the invention provides arrays of diverse double-stranded oligonucleotide sequences. In another embodiment, the invention provides arrays of conformationally restricted probes, wherein the probes are held in position using double-stranded DNA sequences as scaffolding. Libraries of diverse unimolecular double-stranded nucleic acid sequences and probes may be used, for example, in screening studies for determination of binding affinity exhibited by binding proteins, drugs, or RNA.

Methods of synthesizing desired single stranded DNA sequences are well known to those of skill in the art. In particular, methods of synthesizing oligonucleotides are found in, for example, *Oligonucleotide Synthesis: A Practical Approach*, Gait, ed., IRL Press, Oxford (1984), incorporated herein by reference in its entirety for all purposes. Synthesizing unimolecular double-stranded DNA in solution has also been described. See, Durand, et al. *Nucleic Acids Res.* 18:6353–6359 (1990) and Thomson, et al. *Nucleic Acids Res.* 21:5600–5603 (1993), the disclosures of both being incorporated herein by reference.

Solid phase synthesis of biological polymers has been evolving since the early "Merrifield" solid phase peptide synthesis, described in Merrifield, *J. Am. Chem. Soc.* 85:2149–2154 (1963), incorporated herein by reference for all purposes. Solid-phase synthesis techniques have been provided for the synthesis of several peptide sequences on, for example, a number of "pins." See e.g., Geysen et al., *J. Immun. Meth.* 102:259–274 (1987), incorporated herein by reference for all purposes. Other solid-phase techniques involve, for example, synthesis of various peptide sequences on different cellulose disks supported in a column. See Frank and Doring, *Tetrahedron* 44:6031–6040 (1988), incorporated herein by reference for all purposes. Still other solid-phase techniques are described in U.S. Pat. No. 4,728,502 issued to Hamill and WO 90/00626 (Beattie, inventor).

Each of the above techniques produces only a relatively low density array of polymers. For example, the technique described in Geysen et al. is limited to producing 96 different polymers on pins spaced in the dimensions of a standard microtiter plate.

Improved methods of forming large arrays of oligonucleotides, peptides and other polymer sequences in a short period of time have been devised. Of particular note, Pirrung et al., U.S. Pat. No. 5,143,854 (see also PCT Application No. WO 90/15070) and Fodor et al., PCT Publication No. WO 92/10092, all incorporated herein by reference, disclose methods of forming vast arrays of peptides, oligonucleotides and other polymer sequences using, for example, light-directed synthesis techniques. See also, Fodor et al., *Science*, 251:767–777 (1991), also incorporated herein by reference for all purposes. These procedures are now referred to as VLSIPS™ procedures.

In the above-referenced Fodor et al., PCT application, an elegant method is described for using a computer-controlled system to direct a VLSIPS™ procedure. Using this approach, one heterogenous array of polymers is converted, through simultaneous coupling at a number of reaction sites, into a different heterogenous array. See, U.S. Pat. No. 5,384,261 and U.S. application Ser. No. 07/980,523, the disclosures of which are incorporated herein for all purposes.

The development of VLSIPS™ technology as described in the above-noted U.S. Pat. No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, is considered pioneering technology in the fields of combinatorial synthesis and screening of combinatorial libraries. More recently, patent application Ser. No. 08/082,937, filed Jun. 25, 1993 now abandoned, describes methods for making arrays of oligonucleotide probes that can be used to check or determine a partial or complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific oligonucleotide sequence.

A number of biochemical processes of pharmaceutical interest involve the interaction of some species, e.g., a drug, a peptide or protein, or RNA, with double-stranded DNA. For example, protein/DNA binding interactions are involved with a number of transcription factors as well as tumor suppression associated with the p53 protein and the genes contributing to a number of cancer conditions.

SUMMARY OF THE INVENTION

High-density arrays of diverse unimolecular, double-stranded oligonucleotides, as well as arrays of conformationally restricted probes and methods for their use are provided by virtue of the present invention. In addition, methods and devices for detecting duplex formation of oligonucleotides on an array of diverse single-stranded oligonucleotides are also provided by this invention. Further, an adhesive based on the specific binding characteristics of two arrays of complementary oligonucleotides is provided in the present invention.

According to one aspect of the present invention, libraries of unimolecular, double-stranded oligonucleotides are provided. Each member of the library is comprised of a solid support, an optional spacer for attaching the double-stranded oligonucleotide to the support and for providing sufficient space between the double-stranded oligonucleotide and the solid support for subsequent binding studies and assays, an oligonucleotide attached to the spacer and further attached to a second complementary oligonucleotide by means of a flexible linker, such that the two oligonucleotide portions exist in a double-stranded configuration. More particularly, the members of the libraries of the present invention can be represented by the formula:

$$Y—L^1—X^1—L^2—X^2$$

in which Y is a solid support, L^1 is a bond or a spacer, L^2 is a flexible linking group, and X^1 and X^2 are a pair of complementary oligonucleotides.

In a specific aspect of the invention, the library of different unimolecular, double-stranded oligonucleotides can be used for screening a sample for a species which binds to one or more members of the library.

In a related aspect of the invention, a library of different conformationally-restricted probes attached to a solid support is provided. The individual members each have the formula:

$—X^{11}—Z—X^{12}$

in which X^{11} and X^{12} are complementary oligonucleotides and Z is a probe having sufficient length such that X^{11} and X^{12} form a double-stranded oligonucleotide portion of the member and thereby restrict the conformations available to the probe. In a specific aspect of the invention, the library of different conformationally-restricted probes can be used for screening a sample for a species which binds to one or more probes in the library.

According to yet another aspect of the present invention, methods and devices for the bioelectronic detection of duplex formation are provided.

According to still another aspect of the invention, an adhesive is provided which comprises two surfaces of complementary oligonucleotides.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGS. 1A to 1F illustrate the preparation of a member of a library of surface-bound, unimolecular double-stranded DNA as well as binding studies with receptors having specificity for either the double stranded DNA portion, a probe which is held in a conformationally restricted form by DNA scaffolding, or a bulge or loop region of RNA.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Abbreviations

The following abbreviations are used herein: phi, phenanthrenequinone diimine; phen', 5-amido-glutaric acid-1,10-phenanthroline; dppz, dipyridophenazine.

Glossary

The following terms are intended to have the following general meanings as they are used herein:

Chemical terms: As used herein, the term "alkyl" refers to a saturated bydrocarbon radical which may be straight-chain or branched-chain (for example, ethyl, isopropyl, t-amyl, or 2,5-dimethylhexyl). When "alkyl" or "alkylene" is used to refer to a linking group or a spacer, it is taken to be a group having two available valences for covalent attachment, for example, $—CH_2CH_2—$, $—CH_2CH_2CH_2—$, $—CH_2CH_2CH(CH_3)CH_2—$ and $—CH_2(CH_2CH_2)_2CH_2—$. Preferred alkyl groups as substituents are those containing 1 to 10 carbon atoms, with those containing 1 to 6 carbon atoms being particularly preferred. Preferred alkyl or alkylene groups as linking groups are those containing 1 to 20 carbon atoms, with those containing 3 to 6 carbon atoms being particularly preferred. The term "polyethylene glycol" is used to refer to those molecules which have repeating units of ethylene glycol, for example, hexaethylene glycol ($HO—(CH_2CH_2O)_5—CH_2CH_2OH$). When the term "polyethylene glycol" is used to refer to linking groups and spacer groups, it would be understood by one of skill in the art that other polyethers or polyols could be used as well (i. e. polypropylene glycol or mixtures of ethylene and propylene glycols).

The term "protecting group" as used herein, refers to any of the groups which are designed to block one reactive site in a molecule while a chemical reaction is carried out at another reactive site. More particularly, the protecting groups used herein can be any of those groups described in Greene, et al., *Protective Groups In Organic Chemistry*, 2nd Ed., John Wiley & Sons, New York, N.Y., 1991, incorporated herein by reference. The proper selection of protecting groups for a particular synthesis will be governed by the overall methods employed in the synthesis. For example, in "light-directed" synthesis, discussed below, the protecting groups will be photolabile protecting groups such as NVOC, MeNPOC, and those disclosed in co-pending Application PCT/US93/10162 (filed Oct. 22, 1993), incorporated herein by reference. In other methods, protecting groups may be removed by chemical methods and include groups such as FMOC, DMT and others known to those of skill in the art.

Complementary or substantially complementary: Refers to the hybridization or base pairing between nucleotides or nucleic acids, such as, for instance, between the two strands of a double stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single stranded nucleic acid to be sequenced or amplified. Complementary nucleotides are, generally, A and T (or A and U), or C and G. Two single stranded RNA or DNA molecules are said to be substantially complementary when the nucleotides of one strand, optimally aligned and compared and with appropriate nucleotide insertions or deletions, pair with at least about 80% of the nucleotides of the other strand, usually at least about 90% to 95%, and more preferably from about 98 to 100%.

Alternatively, substantial complementary exists when an RNA or DNA strand will hybridize under selective hybridization conditions to its complement. Typically, selective hybridization will occur when there is at least about 65% complementary over a stretch of at least 14 to 25 nucleotides, preferably at least about 75%, more preferably at least about 90% complementary. S. ee, M. Kanehisa *Nucleic Acids Res.* 12:203 (1984), incorporated herein by reference.

Stringent hybridization conditions will typically include salt concentrations of less than about 1M, more usually less than about 500 mM and preferably less than about 200 mM. Hybridization temperatures can be as low as 5° C., but are typically greater than 22° C., more typically greater than about 30° C., and preferably in excess of about 37° C. Longer fragments may require higher hybridization temperatures for specific hybridization. As other factors may affect the stringency, of hybridization, including base composition and length of the complementary strands, presence of organic solvents and extent of base mismatching, the combination of parameters is more important than the absolute measure of any one alone.

Epitope: The portion of an antigen molecule which is delineated by the area of interaction with the subclass of receptors known as antibodies.

Identifier tag: A means whereby one can identify which molecules have experienced a particular reaction in the synthesis of an oligomer. The identifier tag also records the step in the synthesis series in which the molecules experienced that particular monomer reaction. The identifier tag may be any recognizable feature which is, for example: microscopically distinguishable in shape, size, color, optical density, etc.; differently absorbing or emitting of light; chemically reactive; magnetically or electronically encoded; or in some other way distinctively marked with the required information. A preferred example of such an identifier tag is an oligonucleotide sequence.

Ligand/Probe: A ligand is a molecule that is recognized by a particular receptor. The agent bound by or reacting with a receptor is called a "ligand," a term which is definitionally meaningful only in terms of its counterpart receptor. The term "ligand" does not imply any particular molecular size or other structural or compositional feature other than that the substance in question is capable of binding or otherwise interacting with the receptor. Also, a ligand may serve either as the natural ligand to which the receptor binds, or as a functional analogue that may act as an agonist or antagonist.

Examples of ligands that can be investigated by this invention include, but are not restricted to, agonists and antagonists for cell membrane receptors, toxins and venoms, viral epitopes, hormones (e.g., opiates, steroids, etc.), hormone receptors, peptides, enzymes, enzyme substrates, substrate analogs, transition state analogs, cofactors, drugs, proteins, and antibodies. The term "probe" refers to those molecules which are expected to act like ligands but for which binding information is typically unknown. For example, if a receptor is known to bind a ligand which is a peptide β-turn, a "probe" or library of probes will be those molecules designed to mimic the peptide β-turn. In instances where the particular ligand associated with a given receptor is unknown, the term probe refers to those molecules designed as potential ligands for the receptor.

Monomer: Any member of the set of molecules which can be joined together to form an oligomer or polymer. The set of monomers useful in the present invention includes, but is not restricted to, for the example of oligonucleotide synthesis, the set of nucleotides consisting of adenine, thymine, cytosine, guanine, and uridine (A, T, C, G, and U, respectively) and synthetic analogs thereof. As used herein, monomers refers to any member of a basis set for synthesis of an oligomer. Different basis sets of monomers may be used at successive steps in the synthesis of a polymer.

Oligomer or Polymer: The oligomer or polymer sequences of the present invention are formed from the chemical or enzymatic addition of monomer subunits. Such oligomers include, for example, both linear, cyclic, and branched polymers of nucleic acids, polysaccharides, phospholipids, and peptides having either α-, β-, or ω-amino acids, heteropolymers in which a known drug is covalently bound to any of the above, polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneimines, polyarylene sulfides, polysiloxanes, polyimides, polyacetates, or other polymers which will be readily apparent to one skilled in the art upon review of this disclosure. As used herein, the term oligomer or polymer is meant to include such molecules as β-turn mimetics, prostaglandins and benzodiazepines which can also be synthesized in a stepwise fashion on a solid support.

Peptide: A peptide is an oligomer in which the monomers are amino acids and which are joined together through amide bonds and alternatively referred to as a polypeptide. In the context of this specification it should be appreciated that when α-amino acids are used, they may be the L-optical isomer or the D-optical isomer. Other amino acids which are useful in the present invention include unnatural amino acids such a β-alanine, phenylglycine, homoarginine and the like. Peptides are more than two amino acid monomers long, and often more than 20 amino acid monomers long. Standard abbreviations for amino acids are used (e.g., P for proline). These abbreviations are included in Stryer, *Biochemistry*, Third Ed., (1988), which is incorporated herein by reference for all purposes.

Oligonucleotides: An oligonucleotide is a single-stranded DNA or RNA molecule, typically prepared by synthetic means. Alternatively, naturally occurring oligonucleotides, or fragments thereof, may be isolated from their natural sources or purchased from commercial sources. Those oligonucleotides employed in the present invention will be 4 to 100 nucleotides in length, preferably from 6 to 30 nucleotides, although oligonucleotides of different length may be appropriate. Suitable oligonucleotides may be prepared by the phosphoramidite method described by Beaucage and Carruthers, *Tetrahedron Lett.*, 22:1859–1862 (1981), or by the triester method according to Matteucci, et al., *J. Am. Chem. Soc.*, 103:3185 (1981), both incorporated herein by reference, or by other chemical methods using either a commercial automated oligonucleotide synthesizer or VLSIPS™ technology (discussed in detail below). When oligonucleotides are referred to as "double-stranded," it is understood by those of skill in the art that a pair of oligonucleotides exist in a hydrogen-bonded, helical array typically associated with, for example, DNA. In addition to the 100% complementary form of double-stranded oligonucleotides, the term "double-stranded" as used herein is also meant to refer to those forms which include such structural features as bulges and loops, described more fully in such biochemistry texts as Stryer, *Biochemistry*, Third Ed., (1988), previously incorporated herein by reference for all purposes.

Receptor: A molecule that has an affinity for a given ligand or probe. Receptors may be naturally-occurring or manmade molecules. Also, they can be employed in their unaltered natural or isolated state or as aggregates with other species. Receptors may be attached, covalently or noncovalently, to a binding member, either directly or via a specific binding substance. Examples of receptors which can be employed by this invention include, but are not restricted to, antibodies, cell membrane receptors, monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells or other materials), drugs, polynucleotides, nucleic acids, peptides, cofactors, lectins, sugars, polysaccharides, cells, cellular membranes, and organelles. Receptors are sometimes referred to in the art as anti-ligands. As the term receptors is used herein, no difference in meaning is intended. A "ligand-receptor pair" is formed when two molecules have combined through molecular recognition to form a complex. Other examples of receptors which can be investigated by this invention include but are not restricted to:

a) Microorganism receptors: Determination of ligands or probes that bind to receptors, such as specific transport proteins or enzymes essential to survival of microorganisms, is useful in a new class of antibiotics. Of particular value would be antibiotics against opportunistic fungi, protozoa, and those bacteria resistant to the antibiotics in current use.

b) Enzymes: For instance, the binding site of enzymes such as the enzymes responsible for cleaving neurotransmitters. Determination of ligands or probes that bind to certain receptors, and thus modulate the action of the enzymes that cleave the different neurotransmitters, is useful in the development of drugs that can be used in the treatment of disorders of neurotransmission.

c) Antibodies: For instance, the invention may be useful in investigating the ligand-binding site on the antibody molecule which combines with the epitope of an antigen of interest. Determining a sequence that mimics an antigenic epitope may lead to the development of vaccines of which the immunogen is based on one or more of such sequences, or lead to the development of related diagnostic agents or compounds useful in therapeutic treatments such as for autoimmune diseases (e.g., by blocking the binding of the "self" antibodies).

d) Nucleic Acids: The invention may be useful in investigating sequences of nucleic acids acting as binding sites for cellular proteins ("trans-acting factors"). Such sequences may include, e.g., transcription factors, suppressors, enhancers or promoter sequences.

e) Catalytic Polypeptides: Polymers, preferably polypeptides, which are capable of promoting a chemical reaction involving the conversion of one or more reactants to one or more products. Such polypeptides generally include a binding site specific for at least one reactant or reaction intermediate and an active functionality proximate to the binding site, which functionality is capable of chemically modifying the bound reactant. Catalytic polypeptides are described in, Lerner, R.A. et al., *Science* 252: 659 (1991), which is incorporated herein by reference.

f) Hormone receptors: For instance, the receptors for insulin and growth hormone. Determination of the ligands which bind with high affinity to a receptor is useful in the development of, for example, an oral replacement of the daily injections which diabetics must take to relieve the symptoms of diabetes, and in the other case, a replacement for the scarce human growth hormone that can only be obtained from cadavers or by recombinant DNA technology. Other examples are the vasoconstrictive hormone receptors; determination of those ligands that bind to a receptor may lead to the development of drugs to control blood pressure.

g) Opiate receptors: Determination of ligands that bind to the opiate receptors in the brain is useful in the development of less-addictive replacements for morphine and related drugs.

Substrate or Solid Support: A material having a rigid or semi-rigid surface. Such materials will preferably take the form of plates or slides, small beads, pellets, disks or other convenient forms, although other forms may be used. In some embodiments, at least one surface of the substrate will be substantially flat. In other embodiments, a roughly spherical shape is preferred.

Synthetic: Produced by in vitro chemical or enzymatic synthesis. The synthetic libraries of the present invention may be contrasted with those in viral or plasmid vectors, for instance, which may be propagated in bacterial, yeast, or other living hosts.

DESCRIPTION OF THE INVENTION

The broad concept of the present invention is illustrated in FIGS. 1A to 1F. FIGS. 1A, 1B and 1C illustrate the preparation of surface-bound unimolecular double stranded DNA, while FIGS. 1D, 1E, and 1F illustrate uses for the libraries of the present invention.

FIG. 1A shows a solid support 1 having an attached spacer 2, which is optional. Attached to the distal end of the spacer is a first oligomer 3, which can be attached as a single unit or synthesized on the support or spacer in a monomer by monomer approach. FIG. 1B shows a subsequent stage in the preparation of one member of a library according to the present invention. In this stage, a flexible linker 4 is attached to the distal end of the oligomer 3. In other embodiments, the flexible linker will be a probe. FIG. 1C shows the completed surface-bound unimolecular double stranded DNA which is one member of a library, wherein a second oligomer 5 is now attached to the distal end of the flexible linker (or probe). As shown in FIG. 1C, the length of the flexible linker (or probe) 4 is sufficient such that the first and second oligomers (which are complementary) exist in a double-stranded conformation. It will be appreciated by one of skill in the art, that the libraries of the present invention will contain multiple, individually synthesized members which can be screened for various types of activity. Three such binding events are illustrated in FIGS. 1 D, 1E and 1F.

In FIG. 1D, a receptor 6, which can be a protein, RNA molecule or other molecule which is known to bind to DNA, is introduced to the library. Determining which member of a library binds to the receptor provides information which is useful for diagnosing diseases, sequencing DNA or RNA, identifying genetic characteristics, or in drug discovery.

In FIG. 1E, the linker 4 is a probe for which binding information is sought. The probe is held in a conformationally restricted manner by the flanking oligomers 3 and 5, which are present in a double-stranded conformation. As a result, a library of conformationally restricted probes can be screened for binding activity with a receptor 7 which has specificity for the probe.

The present invention also contemplates the preparation of libraries of unimolecular, double-stranded oligonucleotides having bulges or loops in one of the strands as depicted in FIG. 1F. In FIG. 1F, one oligonucleotide 5 is shown as having a bulge 8. Specific RNA bulges are often recognized by proteins (e.g., TAR RNA is recognized by the TAT protein of HIV). Accordingly, libraries of RNA bulges or loops are useful in a number of diagnostic applications. One of skill in the art will appreciate that the bulge or loop can be present in either oligonucleotide portion 3 or 5.

Libraries of Unimolecular, Double-Stranded Oligonucleotides

In one aspect, the present invention provides libraries of unimolecular double-stranded oligonucleotides, each member of the library having the formula:

$$Y—L^1—X^1—L^2—X^2$$

in which Y represents a solid support, X^1 and X^2 represent a pair of complementary oligonucleotides, L^1 represents a bond or a spacer, and L^2 represents a linking group having sufficient length such that X^1 and X^2 form a double-stranded oligonucleotide.

The solid support may be biological, nonbiological, organic, inorganic, or a combination of any of these, existing as particles, strands, precipitates, gels, sheets, tubing, spheres, containers, capillaries, pads, slices, films, plates, slides, etc. The solid support is preferably flat but may take on alternative surface configurations. For example, the solid support may contain raised or depressed regions on which synthesis takes place. In some embodiments, the solid support will be chosen to provide appropriate light-absorbing characteristics. For example, the support may be a polymerized Langmuir Blodgett film, functionalized glass, Si, Ge, GaAs, GaP, SiO_2, SiN_4, modified silicon, or any one of a variety of gels or polymers such as (poly)tetrafluoroethylene, (poly)vinylidendifluoride, polystyrene, polycarbonate, or combinations thereof. Other suitable solid support materials will be readily apparent to those of skill in the art. Preferably, the surface of the solid support will contain reactive groups, which could be carboxyl, amino, hydroxyl, thiol, or the like. More preferably, the surface will be optically transparent and will have surface Si—OH functionalities, such as are found on silica surfaces.

Attached to the solid support is an optional spacer, L^1. The spacer molecules are preferably of sufficient length to permit the double-stranded oligonucleotides in the completed member of the library to interact freely with molecules exposed to the library. The spacer molecules, when present, are typically6–50 atoms long to provide sufficient exposure for the attached double-stranded DNA molecule. The spacer, L^1, is comprised of a surface attaching portion and a longer chain portion. The surface attaching portion is that part of L^1 which is directly attached to the solid support. This portion can be attached to the solid support via carbon-carbon bonds using, for example, supports having (poly)trifluorochloroethylene surfaces, or preferably, by siloxane bonds (using, for example, glass or silicon oxide as the solid support). Siloxane bonds with the surface of the support are formed in one embodiment via reactions of surface attaching portions bearing trichlorosilyl or trialkoxysilyl groups. The surface attaching groups will also have a site for attachment of the longer chain portion. For example, groups which are suitable for attachment to a longer chain portion would include amines, hydroxyl, thiol, and carboxyl. Preferred surface attaching portions include aminoalkylsilanes and hydroxyalkylsilanes. In particularly preferred embodiments, the surface attaching portion of L^1 is either bis(2-hydroxyethyl)-aminopropyltriethoxysilane, 2-hydroxyethylaminopropyltriethoxysilane, aminopropyltriethoxysilane or hydroxypropyltriethoxysilane.

The longer chain portion can be any of a variety of molecules which are inert to the subsequent conditions for polymer synthesis. These longer chain portions will typically be aryl acetylene, ethylene glycol oligomers containing 2–14 monomer units, diamines, diacids, amino acids, peptides, or combinations thereof. In some embodiments, the longer chain portion is a polynucleotide. The longer chain portion which is to be used as part of L^1 can be selected based upon its hydrophilic/hydrophobic properties to improve presentation of the double-stranded oligonucleotides to certain receptors, proteins or drugs. The longer chain portion of L^1 can be constructed of polyethyleneglycols, polynucleotides, alkylene, polyalcohol, polyester, polyamine, polyphosphodiester and combinations thereof. Additionally, for use in synthesis of the libraries of the invention, L^1 will typically have a protecting group, attached to a functional group (i.e., hydroxyl, amino or carboxylic acid) on the distal or terminal end of the chain portion (opposite the solid support). After deprotection and coupling, the distal end is covalently bound to an oligomer.

Attached to the distal end of L^1 is an oligonucleotide, X^1, which is a single-stranded DNA or RNA molecule. The oligonucleotides which are part of the present invention are typically of from about 4 to about 100 nucleotides in length. Preferably, X^1 is an oligonucleotide which is about 6 to about 30 nucleotides in length. The oligonucleotide is typically linked to L^1 via the 3'-hydroxyl group of the oligonucleotide and a functional group on L^1 which results in the formation of an ether, ester, carbamate or phosphate ester linkage.

Attached to the distal end of X^1 is a linking group, L^2, which is flexible and of sufficient length that X^1 can effectively hybridize with X^2. The length of the linker will typically be a length which is at least the length spanned by two nucleotide monomers, and preferably at least four nucleotide monomers, while not be so long as to interfere with either the pairing of X^1 and X^2 or any subsequent assays. The linking group itself will typically be an alkylene group (of from about 6 to about 24 carbons in length), a polyethyleneglycol group (of from about 2 to about 24 ethyleneglycol monomers in a linear configuration), a polyalcohol group, a polyamine group (e.g., spermine, spermidine and polymeric derivatives thereof), a polyester group (e.g., poly(ethyl acrylate) having of from 3 to 15 ethyl acrylate monomers in a linear configuration), a polyphosphodiester group, or a polynucleotide (having from about 2 to about 12 nucleic acids). Preferably, the linking group will be a polyethyleneglycol group which is at least a tetraethyleneglycol, and more preferably, from about 1 to 4 hexaethyleneglycols linked in a linear array. For use in synthesis of the compounds of the invention, the linking group will be provided with functional groups which can be suitably protected or activated. The linking group will be covalently attached to each of the complementary oligonucleotides, X^1 and X^2, by means of an ether, ester, carbamate, phosphate ester or amine linkage. The flexible linking group L^2 will be attached to the 5'-hydroxyl of the terminal monomer of X^1 and to the 3'-hydroxyl of the initial monomer of X^2. Preferred linkages are phosphate ester linkages which can be formed in the same manner as the oligonucleotide linkages which are present in X^1 and X^2. For example, hexaethyleneglycol can be protected on one terminus with a photolabile protecting group (i.e., NVOC or MeNPOC) and activated on the other terminus with 2-cyanoethyl-N,N-diisopropylamino-chlorophosphite to form a phosphoramidite. This linking group can then be used for construction of the libraries in the same manner as the photolabile-protected, phosphoramidite-activated nucleotides. Alternatively, ester linkages to X^1 and X^2 can be formed when the L^2 has terminal carboxylic acid moieties (using the 5'-hydroxyl of X^1 and the 3'-hydroxyl of X^2). Other methods of forming ether, carbamate or amine linkages are known to those of skill in the art and particular reagents and references can be found in such texts as March, *Advanced Organic Chemistry*, 4th Ed., Wiley-Interscience, New York, N.Y. 1992, incorporated herein by reference.

The oligonucleotide, X^2, which is covalently attached to the distal end of the linking group is, like X^1, a single-stranded DNA or RNA molecule. The oligonucleotides which are part of the present invention are typically of from about 4 to about 100 nucleotides in length. Preferably, X^2 is an oligonucleotide which is about 6 to about 30 nucleotides in length and exhibits complementary to X^1 of from 90 to 100%. More preferably, X^1 and X^2 are 100% complementary. In one group of embodiments, either X^1 or X^2 will further comprise a bulge or loop portion and exhibit complementary of from 90 to 100% over the remainder of the oligonucleotide.

In a particularly preferred embodiment, the solid support is a silica support, the spacer is a polyethyleneglycol conjugated to an aminoalkylsilane, the linking group is a polyethyleneglycol group, and X^1 and X^2 are complementary oligonucleotides each comprising of from 6 to 30 nucleic acid monomers.

The library can have virtually any number of different members, and will be limited only by the number or variety of compounds desired to be screened in a given application and by the synthetic capabilities of the practitioner. In one group of embodiments, the library will have from 2 up to 100 members. In other groups of embodiments, the library will have between 100 and 10000 members, and between 10000 and 1000000 members, preferably on a solid support. In preferred embodiments, the library will have a density of more than 100 members at known locations per cm^2, preferably more than 1000 per cm^2, more preferably more than 10,000 per cm^2.

Libraries of Conformationally Restricted Probes

In still another aspect, the present invention provides libraries of conformationally-restricted probes. Each of the members of the library comprises a solid support having an optional spacer which is attached to an oligomer of the formula:

$$-X^{11}-Z-X^{12}$$

in which X^{11} and X^{12} are complementary oligonucleotides and Z is a probe. The probe will have sufficient length such that X^{11} and X^{12} form a double-stranded DNA portion of each member. X^{11} and X^{12} are as described above for X^1 and X^2 respectively, except that for the present aspect of the invention, each member of the probe library can have the same X^{11} and the same X^{12}, and differ only in the probe portion. In one group of embodiments, X^{11} and X^{12} are either a poly-A oligonucleotide or a poly-T oligonucleotide.

As noted above, each member of the library will typically have a different probe portion. The probes, Z, can be any of a variety of structures for which receptor-probe binding information is sought for conformationally-restricted forms. For example, the probe can be an agonist or antagonist for a cell membrane receptor, a toxin, venom, viral epitope, hormone, peptide, enzyme, collector, drug, protein or antibody. In one group of embodiments, the probes are different peptides, each having of from about 4 to about 12 amino acids. Preferably the probes will be linked via polyphosphate diesters, although other linkages are also suitable. For example, the last monomer employed on the X^{11} chain can be a 5'-aminopropyl-functionalized phosphoramidite nucleotide (available from Glen Research, Sterling, Va., USA or Genosys Biotechnologies, The Woodlands, Tex., USA) which will provide a synthesis initiation site for the carboxy to amino synthesis of the peptide probe. Once the peptide probe is formed, a 3'-succinylated nucleoside (from Cruachem, Sterling, Va., USA) will be added under peptide coupling conditions. In yet another group of embodiments, the probes will be oligonucleotides of from 4 to about 30 nucleic acid monomers which will form a DNA or RNA hairpin structure. For use in synthesis, the probes can also have associated functional groups (i.e., hydroxyl, amino, carboxylic acid, anhydride and derivatives thereof) for attaching two positions on the probe to each of the complementary oligonucleotides.

The surface of the solid support is preferably provided with a spacer molecule, although it will be understood that the spacer molecules are not elements of this aspect of the invention. Where present, the spacer molecules will be as described above for L^1.

The libraries of conformationally restricted probes can also have virtually any number of members. As above, the number of members will be limited only by design of the particular screening assay for which the library will be used, and by the synthetic capabilities of the practitioner. In one group of embodiments, the library will have from 2 to 100 members. In other groups of embodiments, the library will have between 100 and 10000 members, and between 10000 and 1000000 members. Also as above, in preferred embodiments, the library will have a density of more than 100 members at known locations per cm^2, preferably more than 1000 per cm^2, more preferably more than 10,000 per cm^2.

Preparation of the Libraries

The present invention further provides methods for the preparation of diverse unimolecular, double-stranded oligonucleotides on a solid support. In one group of embodiments, the surface of a solid support has a plurality of preselected regions. An oligonucleotide of from 6 to 30 monomers is formed on each of the preselected regions. A linking group is then attached to the distal end of each of the oligonucleotides. Finally, a second oligonucleotide is formed on the distal end of each linking group such that the second oligonucleotide is complementary to the oligonucleotide already present in the same preselected region. The linking group used will have sufficient length such that the complementary oligonucleotides form a unimolecular, double-stranded oligonucleotide. In another group of embodiments, each chemically distinct member of the library will be synthesized on a separate solid support.

Libraries on a Single Substrate

Light-Directed Methods

For those embodiments using a single solid support, the oligonucleotides of the present invention can be formed using a variety of techniques known to those skilled in the art of polymer synthesis on solid supports. For example, "light directed" methods (which are one technique in a family of methods known as VLSIPS™ methods) are described in U.S. Pat. No. 5,143,854, previously incorporated by reference. The light directed methods discussed in the '854 patent involve activating predefined regions of a substrate or solid support and then contacting the substrate with a preselected monomer solution. The predefined regions can be activated with a light source, typically shown through a mask (much in the manner of photolithography techniques used in integrated circuit fabrication). Other regions of the substrate remain inactive because they are blocked by the mask from illumination and remain chemically protected. Thus, a light pattern defines which regions of the substrate react with a given monomer. By repeatedly activating different sets of predefined regions and contacting different monomer solutions with the substrate, a diverse array of polymers is produced on the substrate. Of course, other steps such as washing unreacted monomer solution from the substrate can be used as necessary. Other techniques include mechanical techniques such as those described in PCT No. 92/10183, U.S. Pat. No. 5,384,261 also incorporated herein by reference for all purposes. Still further techniques include bead based techniques such as those described in PCT US/93/04145, also incorporated herein by reference, and pin based methods such as those described in U.S. Pat. No. 5,288,514, also incorporated herein by reference.

The VLSIPS™ methods are preferred for making the compounds and libraries of the present invention. The surface of a solid support, optionally modified with spacers having photolabile protecting groups such as NVOC and MeNPOC, is illuminated through a photolithographic mask, yielding reactive groups (typically hydroxyl groups) in the illuminated regions. A 3'-O-phosphoramidite activated deoxynucleoside (protected at the 5'-hydroxyl with a photolabile protecting group) is then presented to the surface and chemical coupling occurs at sites that were exposed to light. Following capping, and oxidation, the substrate is rinsed and the surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of oligonucleotides is produced. Alternatively, an oligomer of from, for example, 4 to 30 nucleotides can be added to each of the preselected regions rather than synthesize each member in a monomer by monomer approach. At this point in the synthesis, either a flexible linking group or a probe can be attached in a similar manner. For example, a flexible linking group such as polyethylene glycol will typically having an activating group (i.e., a phosphoramidite) on one end and a photolabile protecting group attached to the other end. Suitably derivatized polyethylene glycol linking groups can be prepared by the methods described in Durand, et al. *Nucleic Acids Res.* 18:6353–6359 (1990). Briefly, a polyethylene glycol (i.e., hexaethylene glycol) can be mono-protected using MeNPOC-chloride. Following purification of the mono-protected glycol, the remaining hydroxy moiety can be activated with 2-cyanoethyl-N,N-diisopropylaminochlorophosphite. Once the flexible linking group has been attached to the first oligonucleotide (X^1), deprotection and coupling cycles will proceed using 5'-protected, 3'-O-phosphoramidite activated deoxynucleosides or intact oligomers. Probes can be attached in a manner similar to that used for the flexible linking group. When the desired probe is itself an oligomer, it can be formed either in stepwise fashion on the immobilized oligonucleotide or it can be separately synthesized and coupled to the immobilized oligomer in a single step. For example, preparation of conformationally restricted β-turn mimetics will typically involve synthesis of an oligonucleotide as described above, in which the last nucleoside monomer will be derivatized with an aminoalkyl-functionalized phosphoramidite. See, U.S. Pat. No. 5,288,514, previously incorporated by reference. The desired peptide probe is typically formed in the direction from carboxyl to amine terminus. Subsequent coupling of a 3'-succinylated nucleoside, for example, provides the first monomer in the construction of the complementary oligonucleotide strand (which is carried out by the above methods). Alternatively, a library of probes can be prepared by first derivatizing a solid support with multiple poly(A) or poly(T) oligonucleotides which are suitably protected with photolabile protecting groups, deprotecting at known sites and constructing the probe at those sites, then coupling the complementary poly(T) or poly(A) oligonucleotide.

Flow Channel or Spotting Methods

Additional methods applicable to library synthesis on a single substrate are described in co-pending applications Ser. No. 07/980,523, filed Nov. 20, 1992, and U.S. Pat. No. 5,384,261, incorporated herein by reference for all purposes. In the methods disclosed in these applications, reagents are delivered to the substrate by either (1) flowing within a channel defined on predefined regions or (2) "spotting" on predefined regions. However, other approaches, as well as combinations of spotting and flowing, may be employed. In each instance, certain activated regions of the substrate are mechanically separated from other regions when the monomer solutions are delivered to the various reaction sites.

A typical "flow channel" method applied to the compounds and libraries of the present invention can generally be described as follows. Diverse polymer sequences are synthesized at selected regions of a substrate or solid support by forming flow channels on a surface of the substrate through which appropriate reagents flow or in which appropriate reagents are placed. For example, assume a monomer "A" is to be bound to the substrate in a first group of selected regions. If necessary, all or part of the surface of the substrate in all or a part of the selected regions is activated for binding by, for example, flowing appropriate reagents through all or some of the channels, or by washing the entire substrate with appropriate reagents. After placement of a channel block on the surface of the substrate, a reagent having the monomer A flows through or is placed in all or some of the channel(s). The channels provide fluid contact to the first selected regions, thereby binding the monomer A on the substrate directly or indirectly (via a spacer) in the first selected regions.

Thereafter, a monomer B is coupled to second selected regions, some of which may be included among the first selected regions. The second selected regions will be in fluid contact with a second flow channel(s) through translation, rotation, or replacement of the channel block on the surface of the substrate; through opening or closing a selected valve; or through deposition of a layer of chemical or photoresist. If necessary, a step is performed for activating at least the second regions. Thereafter, the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations. In this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, a monomer C can be flowed through still other channels, etc. In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of forming channels or otherwise protecting a portion of the surface of the substrate. For example, according to some embodiments, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further prevented from passing outside of their designated flow paths.

The "spotting" methods of preparing compounds and libraries of the present invention can be implemented in much the same manner as the flow channel methods. For example, a monomer A can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a monomer B can be delivered to and reacted with a second group of activated reaction regions. Unlike the flow channel embodiments described above, reactants are delivered by directly depositing (rather than flowing) relatively small quantities of them in selected regions. In some steps, of course, the entire substrate surface can be sprayed or otherwise coated with a solution. In preferred embodiments, a dispenser moves from region to region, depositing only as much monomer as necessary at each stop. Typical dispensers include a micropipette to deliver the monomer solution to the substrate and a robotic system to control the position of the micropipette with respect to the substrate, or an ink-jet printer. In other embodiments, the dispenser includes a series of tubes, a manifold, an array of pipettes, or the like so that various reagents can be delivered to the reaction regions simultaneously.

Pin-Based Methods

Another method which is useful for the preparation of compounds and libraries of the present invention involves "pin based synthesis." This method is described in detail in U.S. Pat. No. 5,288,514, previously incorporated herein by reference. The method utilizes a substrate having a plurality of pins or other extensions. The pins are each inserted simultaneously into individual reagent containers in a tray. In a common embodiment, an array of 96 pins/containers is utilized.

Each tray is filled with a particular reagent for coupling in a particular chemical reaction on an individual pin. Accordingly, the trays will often contain different reagents. Since the chemistry disclosed herein has been established such that a relatively similar set of reaction conditions may be utilized to perform each of the reactions, it becomes possible to conduct multiple chemical coupling steps simultaneously. In the first step of the process the invention provides for the use of substrate(s) on which the chemical coupling steps are conducted. The substrate is optionally provided with a spacer having active sites. In the particular case of oligonucleotides, for example, the spacer may be selected from a wide variety of molecules which can be used in organic environments associated with synthesis as well as aqueous environments associated with binding studies. Examples of suitable spacers are polyethyleneglycols, dicarboxylic acids, polyamines and alkylenes, substituted with, for example, methoxy and ethoxy groups. Additionally, the spacers will have an active site on the distal end. The active sites are optionally protected initially by protecting groups. Among a wide variety of protecting groups which are useful are FMOC, BOC, t-butyl esters, t-butyl ethers, and the like. Various exemplary protecting groups are described in, for example, Atherton et al., *Solid Phase Peptide Synthesis*, IRL Press (1989), incorporated herein by reference. In some embodiments, the spacer may provide for a cleavable function by way of, for example, exposure to acid or base.

Libraries on Multiple Substrates

Bead Based Methods

Yet another method which is useful for synthesis of compounds and libraries of the present invention involves "bead based synthesis." A general approach for bead based synthesis is described copending application Ser. Nos. 07/762,522 (filed Sep. 18, 1991 now abandoned); 07/946,239 (filed Sep. 16, 1992); 08/146,886 (filed Nov. 2, 1993); 07/876,792 (filed Apr. 29, 1992) and PCT/US93/04145 (filed Apr. 28, 1993), the disclosures of which are incorporated herein by reference.

For the synthesis of molecules such as oligonucleotides on beads, a large plurality of beads are suspended in a suitable carrier (such as water) in a container. The beads are provided with optional spacer molecules having an active site. The active site is protected by an optional protecting group.

In a first step of the synthesis, the beads are divided for coupling into a plurality of containers. For the purposes of this brief description, the number of containers will be limited to three, and the monomers denoted as A, B, C, D, E, and F. The protecting groups are then removed and a first portion of the molecule to be synthesized is added to each of the three containers (i. e., A is added to container 1, B is added to container 2 and C is added to container 3).

Thereafter, the various beads are appropriately washed of excess reagents, and remixed in one container. Again, it will be recognized that by virtue of the large number of beads utilized at the outset, there will similarly be a large number of beads randomly dispersed in the container, each having a particular first portion of the monomer to be synthesized on a surface thereof.

Thereafter, the various beads are again divided for coupling in another group of three containers. The beads in the first container are deprotected and exposed to a second monomer (D), while the beads in the second and third containers are coupled to molecule portions E and F respectively. Accordingly, molecules AD, BD, and CD will be present in the first container, while AE, BE, and CE will be present in the second container, and molecules AF, BF, and CF will be present in the third container. Each bead, however, will have only a single type of molecule on its surface. Thus, all of the possible molecules formed from the first portions A, B, C, and the second portions D, E, and F have been formed.

The beads are then recombined into one container and additional steps such as are conducted to complete the synthesis of the polymer molecules. In a preferred embodiment, the beads are tagged with an identifying tag which is unique to the particular double-stranded oligonucleotide or probe which is present on each bead. A complete description of identifier tags for use in synthetic libraries is provided in co-pending application Ser. No. 08/146,886 (filed Nov. 2, 1993) previously incorporated by reference for all purposes.

Methods of Library Screening

A library prepared according to any of the methods described above can be used to screen for receptors having high affinity for either unimolecular, double-stranded oligonucleotides or conformationally restricted probes. In one group of embodiments, a solution containing a marked (labelled) receptor is introduced to the library and incubated for a suitable period of time. The library is then washed free of unbound receptor and the probes or double-stranded oligonucleotides having high affinity for the receptor are identified by identifying those regions on the surface of the library where markers are located. Suitable markers include, but are not limited to, radiolabels, chromophores, fluorophores, chemiluminescent moieties, and transition metals. Alternatively, the presence of receptors may be detected using a variety of other techniques, such as an assay with a labelled enzyme, antibody, and the like. Other techniques using various marker systems for detecting bound receptor will be readily apparent to those skilled in the art.

In a preferred embodiment, a library prepared on a single solid support (using, for example, the VLSIPS™ technique) can be exposed to a solution containing marked receptor such as a marked antibody. The receptor can be marked in any of a variety of ways, but in one embodiment marking is effected with a radioactive label. The marked antibody binds with high affinity to an immobilized antigen previously localized on the surface. After washing the surface free of unbound receptor, the surface is placed proximate to x-ray film or phosphorimagers to identify the antigens that are recognized by the antibody. Alternatively, a fluorescent marker may be provided and detection may be by way of a charge-coupled device (CCD), fluorescence microscopy or laser scanning.

When autoradiography is the detection method used, the marker is a radioactive label, such as ^{32}P. The marker on the surface is exposed to X-ray film or a phosphorimager, which is developed and read out on a scanner. An exposure time of about 1 hour is typical in one embodiment. Fluorescence detection using a fluorophore label, such as fluorescein, attached to the receptor will usually require shorter exposure times.

Quantitative assays for receptor concentrations can also be performed according to the present invention. In a direct assay method, the surface containing localized probes prepared as described above, is incubated with a solution containing a marked receptor for a suitable period of time. The surface is then washed free of unbound receptor. The amount of marker present at predefined regions of the surface is then measured and can be related to the amount of receptor in solution. Methods and conditions for performing such assays are well-known and are presented in, for example, L. Hood et al., *Immunology*, Benjamin/Cummings (1978), and E. Harlow et al., *Antibodies. A Laboratory Manual*, Cold Spring Harbor Laboratory, (1988). See, also U.S. Pat. No. 4,376,110 for methods of performing sandwich assays. The precise conditions for performing these steps will be apparent to one skilled in the art.

A competitive assay method for two receptors can also be employed using the present invention. Methods of conducting competitive assays are known to those of skill in the art. One such method involves immobilizing conformationally restricted probes on predefined regions of a surface as described above. An unmarked first receptor is then bound to the probes on the surface having a known specific binding affinity for the receptors. A solution containing a marked second receptor is then introduced to the surface and incubated for a suitable time. The surface is then washed free of unbound reagents and the amount of marker remaining on the surface is measured. In another form of competition assay, marked and unmarked receptors can be exposed to the surface simultaneously. The amount of marker remaining on predefined regions of the surface can be related to the amount of unknown receptor in solution. Yet another form of competition assay will utilize two receptors having different labels, for example, two different chromophores.

In other embodiments, in order to detect receptor binding, the double-stranded oligonucleotides which are formed with attached probes or with a flexible linking group will be treated with an intercalating dye, preferably a fluorescent dye. The library can be scanned to establish a background fluorescence. After exposure of the library to a receptor solution, the exposed library will be scanned or illuminated and examined for those areas in which fluorescence has changed. Alternatively, the receptor of interest can be labeled with a fluorescent dye by methods known to those of skill in the art and incubated with the library of probes. The library can then be scanned or illuminated, as above, and examined for areas of fluorescence.

In instances where the libraries are synthesized on beads in a number of containers, the beads are exposed to a receptor of interest. In a preferred embodiment the receptor is fluorescently or radioactively labelled. Thereafter, one or more beads are identified that exhibit significant levels of, for example, fluorescence using one of a variety of techniques. For example, in one embodiment, mechanical separation under a microscope is utilized. The identity of the molecule on the surface of such separated beads is then identified using, for example, NMR, mass spectrometry, PCR amplification and sequencing of the associated DNA, or the like. In another embodiment, automated sorting (i.e., fluorescence activated cell sorting) can be used to separate beads (bearing probes) which bind to receptors from those which do not bind. Typically the beads will be labeled and identified by methods disclosed in Needels, et al., *Proc. Natl. Acad. Sci., USA* 90:10700–10704 (1993), incorporated herein by reference.

The assay methods described above for the libraries of the present invention will have tremendous application in such endeavors as DNA "footprinting" of proteins which bind DNA. Currently, DNA footprinting is conducted using DNase I digestion of double-stranded DNA in the presence of a putative DNA binding protein. Gel analysis of cut and protected DNA fragments then provides a "footprint" of where the protein contacts the DNA. This method is both labor and time intensive. See, Galas et al., *Nucleic Acid Res.* 5:3157 (1978). Using the above methods, a "footprint" could be produced using a single array of unimolecular, double-stranded oligonucleotides in a fraction of the time of conventional methods. Typically, the protein will be labeled with a radioactive or fluorescent species and incubated with a library of unimolecular, double-stranded DNA. Phosphorimaging or fluorescence detection will provide a footprint of those regions on the library where the protein has bound. Alternatively, unlabeled protein can be used. When unlabeled protein is used, the double-stranded oligonucleotides in the library will all be labeled with a marker, typically a fluorescent marker. Incorporation of a marker into each member of the library can be carried out by terminating the oligonucleotide synthesis with a commercially available fluorescing phosphoramidite nucleotide derivative. Following incubation with the unlabeled protein, the library will be treated with DNase I and examined for areas which are protected from cleavage.

The assay methods described above for the libraries of the present invention can also be used in reverse drug discovery. In such an application, a compound having known pharmacological safety or other desired properties (e.g., aspirin) could be screened against a variety of double-stranded oligonucleotides for potential binding. If the compound is shown to bind to a sequence associated with, for example, tumor suppression, the compound can be further examined for efficacy in the related diseases.

In other embodiments, probe arrays comprising β-turn mimetics can be prepared and assayed for activity against a particular receptor. β-turn mimetics are compounds having molecular structures similar to β-turns which are one of the three major components in protein molecular architecture. β-turns are similar in concept to hairpin turns of oligonucleotide strands, and are often critical recognition features for various protein-ligand and protein-protein interactions. As a result, a library of β-turn mimetic probes can provide or suggest new therapeutic agents having a particular affinity for a receptor which will correspond to the affinity exhibited by the β-turn and its receptor.

Bioelectronic Devices and Methods

In another aspect, the present invention provides a method for the bioelectronic detection of sequence-specific oligonucleotide hybridization. A general method and device which is useful in diagnostics in which a biochemical species is attached to the surface of a sensor is described in U.S. Pat. No. 4,562,157 (the Lowe patent), incorporated herein by reference. The present method utilizes arrays of immobilized oligonucleotides (prepared, for example, using VLSIPS™ technology) and the known photo-induced electron transfer which is mediated by a DNA double helix structure. See, Murphy et al., *Science* 262:1025–1029 (1993). This method is useful in hybridizationbased diagnostics, as a replacement for fluorescence-based detection systems. The method of bioelectronic detection also offers higher resolution and potentially higher sensitivity than earlier diagnostic methods involving sequencing/detecting by hybridization. As a result, this method finds applications in genetic mutation screening and primary sequencing of oligonucleotides. The method can also be used for Sequencing By Hybridization (SBH), which is described in copending application Ser. Nos. 08/082,937 (filed Jun. 25, 1993 now abandoned) and 08/168,904 (filed Dec. 15, 1993), each of which are incorporated herein by reference for all purposes. This method uses a set of short oligonucleotide probes of defined sequence to search for complementary sequences on a longer target strand of DNA. The hybridization pattern is used to reconstruct the target DNA sequence. Thus, the hybridization analysis of large numbers of probes can be used to sequence long stretches of DNA. In immediate applications of this hybridization methodology, a small number of probes can be used to interrogate local DNA sequence.

In the present inventive method, hybridization is monitored using bioelectronic detection. In this method, the target DNA, or first oligonucleotide, is provided with an electron-donor tag and then incubated with an array of oligonucleotide probes, each of which bears an electron-acceptor tag and occupies a known position on the surface of the array. After hybridization of the first oligonucleotide to the array has occurred, the hybridized array is illuminated to induce an electron transfer reaction in the direction of the surface of the array. The electron transfer reaction is then detected at the location on the surface where hybridization has taken place. Typically, each of the oligonucleotide probes in an array will have an attached electron-acceptor tag located near the surface of the solid support used in preparation of the array. In embodiments in which the arrays are prepared by light-directed methods (i.e. typically 3' to 5' direction), the electronacceptor tag will be located near the 3' position. The electron-acceptor tag can be attached either to the 3' monomer by methods known to those of skill in the art, or it can be attached to a spacing group between the 3' monomer and the solid support. Such a spacing group will have, in addition to functional groups for attachment to the solid support and the oligonucleotide, a third functional group for attachment of the electronacceptor tag. The target oligonucleotide will typically have the electron-donor tag attached at the 3' position. Alternatively, the target oligonucleotide can be incubated with the array in the absence of an electron-donor tag. Following incubation, the electron-donor tag can be added in solution. The electron-donor tag will then intercalate into those regions where hybridization has occurred. An electron transfer reaction can then be detected in those regions having a continuous DNA double helix.

The electron-donor tag can be any of a variety of complexes which participate in electron transfer reactions and which can be attached to an oligonucleotide by a means which does not interfere with the electron transfer reaction. In preferred embodiments, the electron-donor tag is a ruthenium (II) complex, more preferably a ruthenium (II) (phen')$_2$(dppz) complex.

The electron-acceptor tag can be any species which, with the electron-donor tag, will participate in an electron transfer reaction. An example of an electron-acceptor tag is a rhodium (III) complex. A preferred electron-acceptor tag is a rhodium (III) (phi)$_2$(phen') complex.

In a particularly preferred embodiment, the electron-donor tag is a ruthenium (II) (phen')$_2$(dppz) complex and the electron-acceptor tag is a rhodium (III) (phi)$_2$(phen') complex.

In still another aspect, the present invention provides a device for the bioelectronic detection of sequence-specific oligonucleotide hybridization. The device will typically consist of a sensor having a surface to which an array of oligonucleotides are attached. The oligonucleotides will be attached in pre-defined areas on the surface of the sensor and have an electron-acceptor tag attached to each oligonucleotide. The electron-acceptor tag will be a tag which is capable of producing an electron transfer signal upon illumination of a hybridized species, when the complementary oligonucleotide bears an electrondonating tag. The signal will be in the direction of the sensor surface and be detected by the sensor.

In a preferred embodiment, the sensor surface will be a silicon-based surface which can sense the electronic signal induced and, if necessary, amplify the signal. The metal contacts on which the probes will be synthesized can be treated with an oxygen plasma prior to synthesis of the probes to enhance the silane adhesion and concentration on the surface. The surface will further comprise a multi-gated field effect transistor, with each gate serving as a sensor and different oligonucleotides attached to each gate. The oligonucleotides will typically be attached to the metal contacts on the sensor surface by means of a spacer group.

The spacer group should not be too long, in order to ensure that the sensing function of the device is easily activated by the binding interaction and subsequent illumination of the "tagged" hybridized oligonucleotides. Preferably, the spacer group is from 3 to 12 atoms in length and will be as described above for the surface modifying portion of the spacer group, L^1.

The oligonucleotides which are attached to the spacer group can be formed by any of the solid phase techniques which are known to those of skill in the art. Preferably, the oligonucleotides are formed one base at a time in the direction of the 3' terminus to the 5' terminus by the "light-directed" methods described above. The oligonucleotide can then be modified at the 3' end to attach the electron-acceptor tag. A number of suitable methods of attachment are known. For example, modification with the reagent Aminolink2 (from Applied Biosystems, Inc.) provides a terminal phosphate moiety which is derivatized with an aminohexyl phosphate ester. Coupling of a carboxylic acid, which is present on the electron-acceptor tag, to the amine can then be carried out using HOBT and DCC. Alternatively, synthesis of the oligonucleotide can begin with a suitably derivatized and protected monomer which can then be deprotected and coupled to the electron-acceptor tag once the complete oligonucleotide has been synthesized.

The silica surface can also be replaced by silicon nitride or oxynitride, or by an oxide of another metal, especially aluminum, titanium (IV) or iron (III). The surface can also be any other film, membrane, insulator or semiconductor overlying the sensor which will not interfere with the detection of electron transfer detection and to which an oligonucleotide can be coupled.

Additionally, detection devices other than an FET can be used. For example, sensors such as bipolar transistors, MOS transistors and the like are also useful for the detection of electron transfer signals.

Adhesives

In still another aspect, the present invention provides an adhesive comprising a pair of surfaces, each having a plurality of attached oligonucleotides, wherein the single-stranded oligonucleotides on one surface are complementary to the single-stranded oligonucleotides on the other surface. The strength and position/orientation specificity can be controlled using a number of factors including the number and length of oligonucleotides on each surface, the degree of complementary, and the spatial arrangement of complementary oligonucleotides on the surface. For example, increasing the number and length of the oligonucleotides on each surface will provide a stronger adhesive. Suitable lengths of oligonucleotides are typically from about 10 to about 70 nucleotides. Additionally, the surfaces of oligonucleotides can be prepared such that adhesion occurs in an extremely position-specific manner by a suitable arrangement of complementary oligonucleotides in a specific pattern. Small deviations from the optimum spatial arrangement are energetically unfavorable as many hybridization bonds must be broken and are not reformed in any other relative orientation.

The adhesives of the present invention will find use in numerous applications. Generally, the adhesives are useful for adhering two surfaces to one another. More specifically, the adhesives will find application where biological compatibility of the adhesive is desired. An example of a biological application involves use in surgical procedures where tissues must be held in fixed positions during or following the procedure. In this application, the surfaces of the adhesive will typically be membranes which are compatible with the tissues to which they are attached.

A particular advantage of the adhesives of the present invention is that when they are formed in an orientation specific manner, the adhesive portions will be "self-finding," that is the system will go to the thermodynamic equilibrium in which the two sides are matched in the predetermined, orientation specific manner.

EXAMPLES

Example 1

This example illustrates the general synthesis of an array of unimolecular, double-stranded oligonucleotides on a solid support.

Unimolecular double stranded DNA molecules were synthesized on a solid support using standard light-directed methods (VLSIPS™ protocols). Two hexaethylene glycol (PEG) linkers were used to covalently attach the synthesized oligonucleotides to the derivatized glass surface. Synthesis of the first (inner) strand proceeded one nucleotide at a time using repeated cycles of photo-deprotection and chemical coupling of protected nucleotides. The nucleotides each had a protecting group on the base portion of the monomer as well as a photolabile MeNPoc protecting group on the 5' hydroxyl. Upon completion of the inner strand, another MeNPoc-protected PEG linker was covalently attached to the 5' end of the surface-bound oligonucleotide. After addition of the internal PEG linker, the PEG is photodeprotected, and the synthesis of the second strand proceeded in the normal fashion. Following the synthesis cycles, the DNA bases were deprotected using standard protocols. The sequence of the second (outer) strand, being complementary to that of the inner strand, provided molecules with short, hydrogen bonded, unimolecular double-stranded structure as a result of the presence of the internal flexible PEG linker.

An array of 16 different molecules were synthesized on a derivatized glass slide in order to determine whether short, unimolecular DNA structures could be formed on a surface and whether they could adopt structures that are recognized by proteins. Each of the 16 different molecular species occupies a different physical region on the glass surface so that there is a one-to-one correspondence between molecular identity and physical location. The molecules are of the form

S-P-P-C-C-A/T-A/T-A/T-A/T-G-C-P-G-C-A/T-A/T-A/T-A/T-G-G-F

where S is the solid surface having silyl groups, P is a PEG linker, A, C, G, and T are the DNA nucleotides, and F is a fluorescent tag. The DNA sequence is listed from the 3' to the 5' end (the 3' end of the DNA molecule is attached to the solid surface via a silyl group and 2 PEG linkers). The sixteen molecules synthesized on the solid support differed in the various permutations of A and T in the above formula.

Example 2

This example illustrates the ability of a library of surface-bound, unimolecular, double-stranded oligonucleotides to exist in duplex form and to be recognized and bound by a protein.

A library of 16 different members was prepared as described in Example 1. The 16 molecules all have the same composition (same number of As, Cs, Gs and Ts), but the order is different. Four of the molecules have an outer strand that is 100% complementary to the inner strand (these molecules will be referred to as DS, doublestranded, below). One of the four DS oligonucleotides has a sequence that is recognized by the restriction enzyme EcoR1. If the molecule can loop back and form a DNA duplex, it should be recognized and cut by the restriction enzyme, thereby releasing the fluorescent tag. Thus, the action of the enzyme provided a functional test for DNA structure, and also served to demonstrate that these structures can be recognized at the surface by proteins. The remaining 12 molecules had outer strands that were not complementary to their inner strands (referred to as SS, single-stranded, below). Of these, three had an outer strand and three had an inner strand whose sequence was an EcoR1 half-site (the sequence on one strand was correct for the enzyme, but the other half was not). The solid support with an array of molecules on the surface is referred to as a "chip" for the purposes of the following discussion. The presence of fluorescently labelled molecules on the chip was detected using confocal fluorescence microscopy. The action of various enzymes was determined by monitoring the change in the amount of fluorescence from the molecules on the chip surface (e.g. "reading" the chip) upon treatment with enzymes that can cut the DNA and release the fluorescent tag at the 5' end.

The three different enzymes used to characterize the structure of the molecules on the chip were:

1) Mung Bean Nuclease—sequence independent, single-strand specific DNA endonuclease;

2) DNase I—sequence independent, double-strand specific endonuclease;

3) EcoR1—restriction endonuclease that recognizes the sequence (5'-3')

GAATTC in double stranded DNA, and cuts between the G and the first A. Mung Bean Nuclease and EcoR1 were obtained from New England Biolabs, and DNase I was obtained from Boehringer Mannheim. All enzymes were used at a concentration of 200 units per mL in the buffer recommended by the manufacturer. The enzymatic reactions were performed in a 1 mL flow cell at 22° C., and were typically allowed to proceed for 90 minutes.

Upon treatment of the chip with the enzyme EcoR1, the fluorescence signal in the DS EcoR1 region and the 3 SS regions with the EcoR1 half-site on the outer strand was reduced by about 10% of its initial value. This reduction was at least 5 times greater than for the other regions of the chip, indicating that the action of the enzyme is sequence specific on the chip. It was not possible to determine if the factor is greater than 5 in these preliminary experiments because of uncertainty in the constancy of the fluorescence background. However, because the purpose of these early experiments was to determine whether unimolecular double-stranded structures could be formed and whether they could be specifically recognized by proteins (and not to provide a quantitative measure of enzyme specificity), qualitative differences between the different synthesis regions were sufficient.

The reduction in signal in the 3 SS regions with the EcoR1 half-site on the outer strand indicated either that the enzyme cuts single-stranded DNA with a particular sequence, or that these molecules formed a double-stranded structure that was recognized by the enzyme. The molecules on the chip surface were at a relatively high density, with an average spacing of approximately 100 angstroms. Thus, it was possible for the outer strand of one molecule to form a double-stranded structure with the outer strand of a neighboring molecule. In the case of the 3 SS regions with the EcoR1 half-site on the outer strand, such a bimolecular double-stranded region would have the correct sequence and structure to be recognized by EcoR1. However, it would differ from the unimolecular double-stranded molecules in that the inner strand remains single-stranded and thus amenable to cleavage by a single-strand specific endonuclease such as Mung Bean Nuclease. Therefore, it was possible to distinguish unimolecular from bimolecular double-stranded

DNA molecules on the surface by their ability to be cut by single and double-strand specific endonucleases.

In order to remove all molecules that have single-stranded structures and to identify unimolecular double-stranded molecules, the chip was first exhaustively treated with Mung Bean Nuclease. The reduction in the fluorescence signal was greater by about a factor of 2 for the SS regions of the chip, including those with the EcoR1 half-site on the outer strand that were cleaved by EcoR1, than for the 4 DS regions. Following Mung Bean Nuclease treatment, the chip was treated with either DNase I (which cuts all remaining double-stranded molecules) or EcoR1 (which should cut only the remaining double-stranded molecules with the correct sequence). Upon treatment with DNase I, the fluorescence signal in the 4 DS regions was reduced by at least 5-fold more than the signal in the SS regions. Upon EcoR1 treatment, the signal in the single DS region with the correct EcoR1 sequence was reduced by at least a factor of 3 more than the signal in any other region on the chip. Taken together, these results indicated that the surface-bound molecules synthesized with two complementary strands separated by a flexible PEG linker form intramolecular double-stranded structures that were resistant to a single-strand specific endonuclease and were recognized by both a double-strand specific endonuclease, and a sequence-specific restriction enzyme.

What is claimed is:

1. A synthetic unimolecular, double-stranded oligonucleotide library comprising a plurality of different members, each member having the formula:

$$Y—L^1—X^1—L^2—X^2$$

wherein,

Y is a solid support;

X^1 and X^2 are a pair of complementary oligonucleotides;

L^1 is a spacer;

L^2 is a linking group having sufficient length such that X^1 and X^2 form a double-stranded oligonucleotide.

2. A library in accordance with claim 1, wherein L^2 is a polyethylene glycol group.

3. A library in accordance with claim 1, wherein X^1 and X^2 are complementary oligonucleotides each comprising of from 6 to 30 nucleic acid monomers.

4. A library in accordance with claim 1, wherein said solid support is a silica support and L^1 comprises an aminoalkylsilane and from 1 to 4 hexaethyleneglycols.

5. A library in accordance with claim 1, wherein said solid support is a silica support, L^1 comprises an aminoalkylsilane and from 1 to 4 hexaethyleneglycols, L^2 is a polyethyleneglycol group and X^1 and X^2 are complementary oligonucleotides each comprising of from 6 to 30 nucleic acid monomers.

6. A synthetic unimolecular, double-stranded oligonucleotide library of claim 1, wherein a portion of said double-stranded oligonucleotides formed by X^1 and X^2 further comprise a loop.

* * * * *

Y

US005744305A

United States Patent [19]

Fodor et al.

[11] Patent Number: 5,744,305

[45] Date of Patent: *Apr. 28, 1998

[54] ARRAYS OF MATERIALS ATTACHED TO A SUBSTRATE

[75] Inventors: Stephen P.A. Fodor, Palo Alto; Lubert Stryer, Stanford; J. Leighton Read, Palo Alto, all of Calif.; Michael C. Pirrung, Durham, N.C.

[73] Assignee: Affymetrix, Inc., Santa Clara, Calif.

[*] Notice: The term of this patent shall not extend beyond the expiration date of Pat. No. 5,445,934.

[21] Appl. No.: 466,632

[22] Filed: Jun. 6, 1995

Related U.S. Application Data

[60] Division of Ser. No. 390,272, Feb. 16, 1995, Pat. No. 5,489,678, and a continuation-in-part of Ser. No. 456,887, Jun. 1, 1995, which is a division of Ser. No. 954,646, Sep. 30, 1992, Pat. No. 5,445,934, which is a division of Ser. No. 850,356, Mar. 12, 1992, Pat. No. 5,405,783, which is a division of Ser. No. 492,462, Mar. 7, 1990, Pat. No. 5,143,854, which is a continuation-in-part of Ser. No. 362,901, Jun. 7, 1989, abandoned, said Ser. No. 390,272, is a continuation of Ser. No. 624,120, Dec. 6, 1990, abandoned, which is a continuation-in-part of Ser. No. 492,462, Mar. 7, 1990, Pat. No. 5,143,854, which is a continuation-in-part of Ser. No. 362,901, Jun. 7, 1989, abandoned.

[51] Int. Cl.^{6} C12Q 1/68; C07H 21/04; C07H 21/02

[52] U.S. Cl. 435/6; 435/7.92; 435/7.94; 435/7.95; 435/969; 435/973; 436/518; 436/527; 436/807; 436/809; 530/334; 536/24.3; 536/25.3; 536/25.32

[58] Field of Search 435/6, 7.92, 7.95, 435/7.94, 9.73, 969; 436/518, 527, 807, 809; 530/334; 536/24.3, 25.3, 25.32

[56] References Cited

U.S. PATENT DOCUMENTS

3,849,137	11/1974	Barzynski et al. .	
4,269,933	5/1981	Pazos .	
4,516,833	5/1985	Pusek .	
4,517,338	5/1985	Urdea et al. .	
4,537,861	8/1985	Elings et al. .	
4,562,157	12/1985	Lowe et al.	435/291
4,631,211	12/1986	Houghton	428/35
4,689,405	8/1987	Frank et al. .	
4,704,353	11/1987	Humphries et al.	435/4
4,713,326	12/1987	Dattagupta et al.	435/6
4,728,591	3/1988	Clark et al. .	
4,762,881	8/1988	Kauer	525/54.11

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

0 046 083	2/1982	European Pat. Off.	B01F 17/00
0 103 197	3/1984	European Pat. Off.	C07C 79/46
0 228 310	10/1988	European Pat. Off. .	
0 328 256 A1	1/1989	European Pat. Off.	B01J 20/32
0 392 546	10/1990	European Pat. Off.	C12Q 1/68

(List continued on next page.)

OTHER PUBLICATIONS

Barinaga, "Will 'DNA Chip' Spped Genome Initiative?" *Science*, 253:1489 (1991).

BioRad Çatalogue M 1987, p. 182.

Geyson et al., "Strategies for epitope analysis using peptide synthesis," *J. Immunol. Methods*, 102:259–274 (1987).

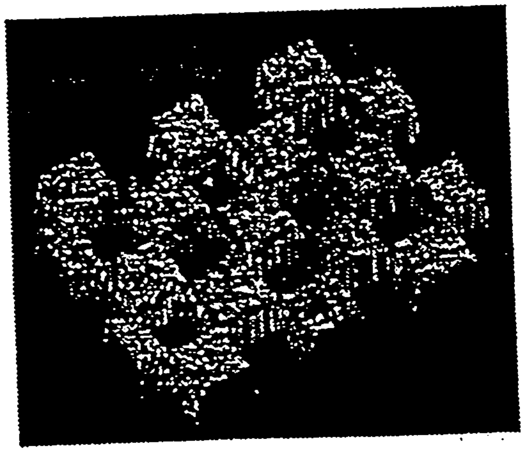
(List continued on next page.)

Primary Examiner—Stephanie W. Zitomer
Assistant Examiner—Paul B. Tran
Attorney, Agent, or Firm—Vern Norviel; Nancy J. DeSantis; Joseph Liebeschuetz

[57] ABSTRACT

A synthetic strategy for the creation of large scale chemical diversity. Solid-phase chemistry, photolabile protecting groups, and photolithography are used to achieve light-directed spatially-addressable parallel chemical synthesis. Binary masking techniques are utilized in one embodiment. A reactor system, photoremovable protective groups, and improved data collection and handling techniques are also disclosed. A technique for screening linker molecules is also provided.

26 Claims, 17 Drawing Sheets



U.S. PATENT DOCUMENTS

4,811,062	3/1989	Taba et al. .	
4,833,092	5/1989	Geysen	436/501
4,846,552	7/1989	Veldkamp et al.	350/162.2
4,886,741	12/1989	Schwartz .	
4,888,278	12/1989	Singer et al. .	
4,923,901	5/1990	Koester et al. .	
4,946,942	8/1990	Fuller et al.	530/335
4,973,493	11/1990	Guire .	
4,981,985	1/1991	Kaplan et al. .	
4,984,100	1/1991	Takayama et al.	360/49
5,079,600	1/1992	Schnur et al. .	
5,143,854	9/1992	Pirrung et al. .	
5,202,231	4/1993	Drmanac et al. .	
5,252,743	10/1993	Barrett et al. .	
5,258,506	11/1993	Urdea et al. .	
5,445,934	8/1995	Fodor et al.	435/6

FOREIGN PATENT DOCUMENTS

1-233 447	9/1989	Japan .	
2 196 476	2/1990	United Kingdom	H01L 21/46
WO 84/03564	9/1984	WIPO	G01N 33/54
WO 86/06487	11/1986	WIPO	G01N 33/53
WO 89/10977	11/1989	WIPO	C12Q 1/68
WO 89/11548	11/1989	WIPO	C12Q 1/68
WO 90/03382	4/1990	WIPO	C07H 21/00
WO 90/04652	5/1990	WIPO	C12Q 1/68
WO 90/15070	12/1990	WIPO	C07K 1/104
WO 91/07087	5/1991	WIPO	A01N 1/02

OTHER PUBLICATIONS

Haynes & Higgens (eds.), *Nucleic Acid Hybridization: A Practical Approach*, IRL Press, Oxford, England, pp. 126–128 (1985).

Mirzabekov, "DNA sequencing by hybridization –a megasequencing method and a diagnostic tool?" *TIBTECH*, 12:27–32 (1994).

Southern et al., "Analyzing and Comparing Nucleic Acid Sequences by Hybridization to Arrays of oligonucleotides: Evaluation Using Experimental Models," *Genomics*, 13:1008–1017 (1992).

"A Sequencing Reality Check", *Research News* [*Science*] (1988) 242:1245.

"Affymax Raises $25 Million to Develop High Speed Drug Discovery System" *Biotechnology News*, vol. 10, No. 3, Feb. 1, 1990, pp. 7–8.

Adams et al., "Photolabile chelators that 'cage' calcium with improved speed of release and pre-photolysis affinity," *J. General Physiology* (Dec. 1986).

Adams et al., "Biologically useful chelators that take up Ca^{2+} upon illumination," *J. Am. Chem. Soc.* 111:7957–7968 (1989).

Amit et al., "Photosensitive protecting groups of amino sugars and their use in glycoside synthesis. 2-Nitrobenzy-loxycarbonylamino and 6-Nitroveratryloxycarbonylamino derivatives," *J. Org. Chem.* (1974) 39:192–196.

Amit et al., "Photosensitive protecting groups –A review," *Israel J. of Chem.* 12(1–2):103–113 (1974).

Bains et al., "A novel method for nucleic acid sequence determination," *J. Theor. Biol.* (1988) 135:303–307.

Baldwin et al., "New photolabile phosphate protecting groups," *Tetrahedron* 46(19):6879–6884 (1990).

Barltrop et al., "Photosensitive protective groups," *Chemical Communications*, p. 822 (Nov. 22, 1966).

Cameron et al., "Photogeneration of organic bases from o-nitrobenzyl-derived carbamates," *J. Am. Chem. Soc.* (1991) 113, 4303–4313.

Craig et al., "Ordering of cosmid clones covering the herpes simplex virus type 1 (HSV-1) genome: a test case for fingerprinting by hybridisation," *Nucl. Acids Res.* (1990) 18:2653–2660.

Cummings et al., "Photoactivable fluorophores. 1. Synthesis and photoactivation of o-nitrobenzyl-quenched fluorescent carbamates," *Tetrahedron Letters* (1988) 29:65–68.

Drmanac et al., "Sequencing of megabase plus DNA by hybridization: theory of the method," *Genomics* (1989) 4:114–128.

Dulcey et al., "Deep UV photochemistry of chemisorbed monolayers: patterned coplanar molecular assemblies," *Science* (1991) 252:551–554.

Flanders et al., "A new interferometric alignment technique," *App. Phys. Lett.* (1977) 31:426–428.

Fodor et al., "Light-directed Spatially-addressable Parallel Chemical Synthesis," *Science* 251:767–773 (1991).

Furka et al., "General method for rapid synthesis of multicomponent peptide mixtures," *Int. J. Peptide Protein Res.* (1991) 37:487–493.

Furka, et al., "Cornucopia of peptides by synthesis," *14th Int'l Congress of Biochem.*, Abstract No. FR:013, Prague, Czechoslovakia, Jul. 10–15, 1988.

Furka et al., "More peptides by less labour," *Xth Int'l Symposium on Medicinal Chemistry*, (Abstract No. 288) Budapest, Hungary, Aug. 15–19, 1988.

Gurney et al., Activation of a potassium current by rapid photochemically generated step increases of intracellular calcium in rat sympathetic neurons, *PNAS USA* 84:3496–3500 (May 1987).

Haridasan et al., "Peptide synthesis using photolytically cleavable 2-nitrobenzyloxycarbonyl protecting group," *Proc. Indian Natl. Sci. Acad.* Part A (1987) 53:717–728.

Iwamura et al., "1-pyrenylmethyl esters, photolabile protecting groups for carboxylic acids," *Tetrahedron Letters* (1987) 28:679–682.

Iwamura et al., "1-x-Diazobenyl pyrene: A reagent for photolabile and fluorescent protection of carboxyl groups of amino acids and peptides," *Chemical Abstracts*, vol. 114(23) (1991).

Iwamura et al., "1-(α-Diazobenyl pyrene: A reagent for photolabile and fluorescent protection of carboxyl groups of amino acids and peptides," (1991) *Synlett* 35–36.

Kaplan et al., "Photolabile chelators for the rapid photorelease of divalent cations," *PNAS USA* 85:6571–6575 (Sep. 1988).

Khrapko et al., "An oligonucleotide hybridization approach to DNA sequencing," *FEBS. Lett.* (1989) 256:118–122.

Kleinfeld et al., "Controlled outgrowth of dissociated neurons on patterned substrates," *J. of Neuroscience* 8(11):4098–4120 (Nov. 1988).

Krile et al., "Multiplex holography with chrip-modulated binary phase-coded reference-beam masks," *Applied Optics* (1979) 18:52–56.

Lam et al., "A new type of synthetic peptide library for identifying ligand-binding activity," *Nature* (1991) 354:82–86.

Logue et al., "General approaches to mask design for binary optics," *SPIE* (1989) 1052:19–24.

Lysov et al., "A new method for determining the DNA nucleotide sequence by hybridization with oligonucleotides," *Doklady Akademii Nauk SSR* (1988) 303:1508–1511.

McCray et al., "Properties and uses of photoreactive caged compounds," *Ann. Rev. Biophys. and Biophys. Chem.* (1989) 18:239–270.

McGillis, "Lithography," *VLSI Technology*, S. Sze, ed., McGraw–Hill Book Company, 1983, pp. 267–301.

Ohtsuka et al., "Studies on transfer ribonucleic acids and related compounds. IX(1) Ribooligonucleotide synthesis using a photosensitive o–nitrobenzyl protection at the 2'–hydroxyl group," *Nucleic Acids Research* (1974) 1:1351–1357.

Patchornik et al., "Photosensitive protecting groups," *J. Am. Chem. Soc.* (1970) 92:6333–6335.

Patent Abstracts of Japan from the EPO, Abst. vol. 13:557, pub. date 12–28–89 abstracting Japanese Patent 01–233 447.

Pillai et al., "3–nitro–4–aminomethyl–benzoylderivate von poly–ethylenglykolen: eine neue klasse von photosensitiven loslichen polymeren tragern zur synthese von c–terminalen peptidamiden." *Tetrahedron Letters* (1979) No. 36, pp. 3409–3412.

Pillai et al., "Photoremovable protecting groups in organic synthesis," *Synthesis* pp. 1–26 (Jan. 1980).

Poustka et al., "Molecular approaches to mammalian genetics," *CSH Symp. Quant. Biol.* (1986) 51:131–139.

Reichmanis et al., "o–nitrobenzyl photochemistry: Solution vs. solid–state behavior," *J. Polymer Sc. Polymer Chem. Ed.* 23:1–8 (1985).

Robertson et al., "A general and efficient route for chemical aminoacylation of transfer RNAs," *Chemical Abstracts*, vol. 114, No. 15 (1991).

Robertson et al., "A general and efficient route for chemical aminoacylation of transfer RNAs," (1991) *J. Am. chem. Soc.* 113:2722–2729.

Schuup et al., "Mechanistic studies of the photorearrangement of o–nitrobenzyl esters," *J. Photochem.* 36:85–97 (1987).

Shin et al., "Dehydrooligopeptides. XI. Facile syntheses of various kinds of dehydrodi–and tripeptides, and dehydroenkephalins containing Δtyr residence by using N–carboxydehydrotyrosine anhydride," (1989) *Bull Chem. Soc. Jpn.* 62:1127–1135.

Shin et al., "Dehydrooligopeptides. XI. Facile synthesis of various kinds of dehydrodi–and tripeptides, and dehydroenkephalins containing tyr residence by using N–carboxydehydrotyrosine anhydride," *Chemical Abstracts*, 112(11) 1990).

Tsien et al., "Control of cytoplasmic calcium with photolabile tetracarboxylate 2–nitrobenzyhydrol chelators," *Biophys. J.*50:843–853 (Nov. 1986).

Veldkamp, "Binary optics: the optics technology of the 1990s," *CLEO 90*, May 21, 1990, Paper No. CMG6.

Walker et al., "Photolabile protecting groups for an acetyl–choline receptor ligand. Synthesis and photochemistry of a new class of o–nitrobenzyl derivatives and their effects on receptor function," *Biochemistry* 25:1799–1805 (1986).

Zehavi et al., "Light–sensitive glycosides. I. 6–nitroveratryl β–D–glucopyranoside and 2–nitrobenzyl β–D–glucopyranoside," *J. Org. Chem.* (1972) 37:2281–2285.

Wilcox et al., "Synthesis of photolabile 'precursors' of amino acid neurotransmitters," *J. Org. Chem.* 55:1585–1589 (1990).

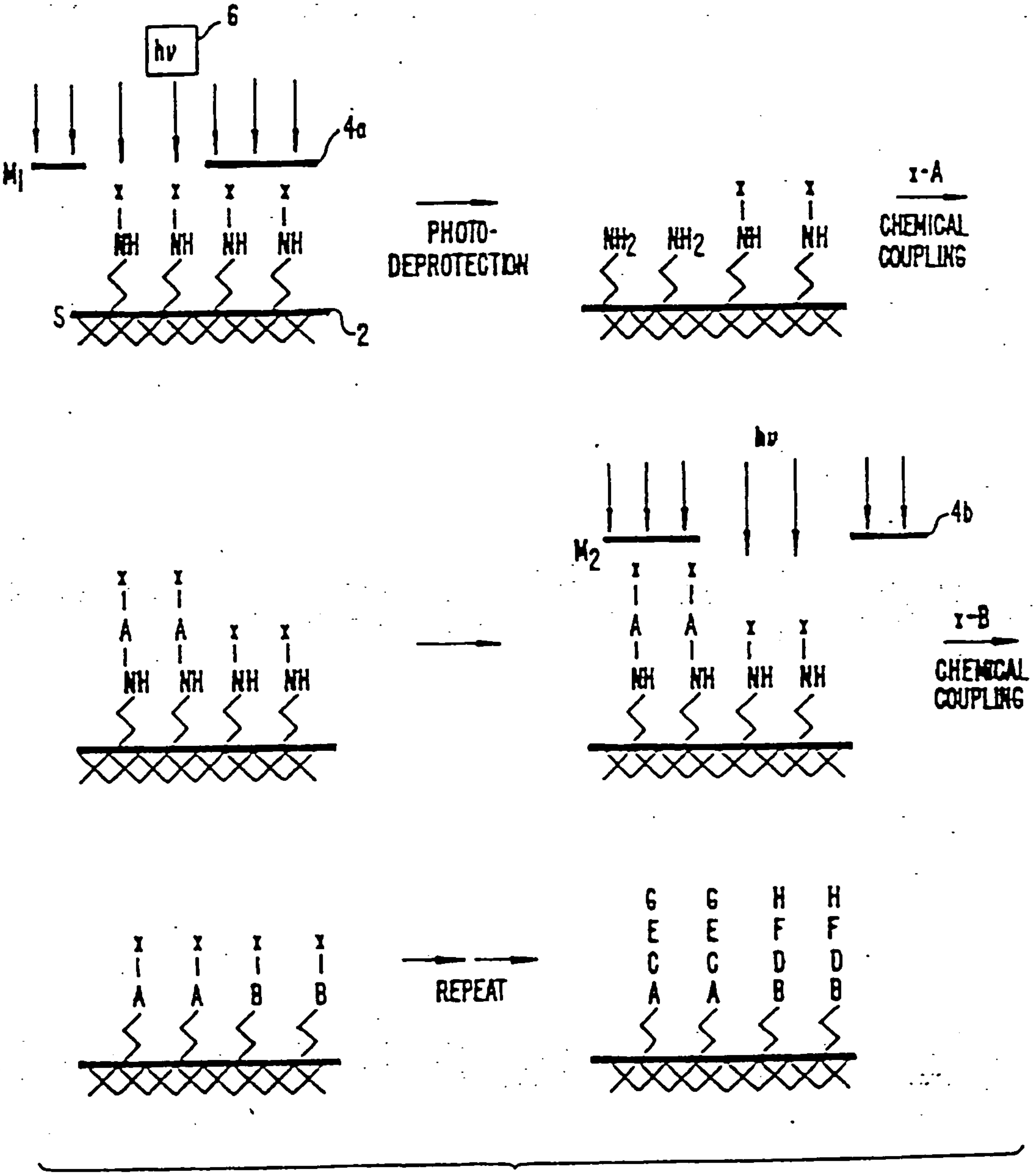

hν
6
4a
M1
X
NH
S
2
PHOTO-DEPROTECTION
NH2
X-A
CHEMICAL COUPLING
hν
4b
M2
A
X-B
CHEMICAL COUPLING
B
REPEAT
G
E
C
H
F
D

FIG. 1.

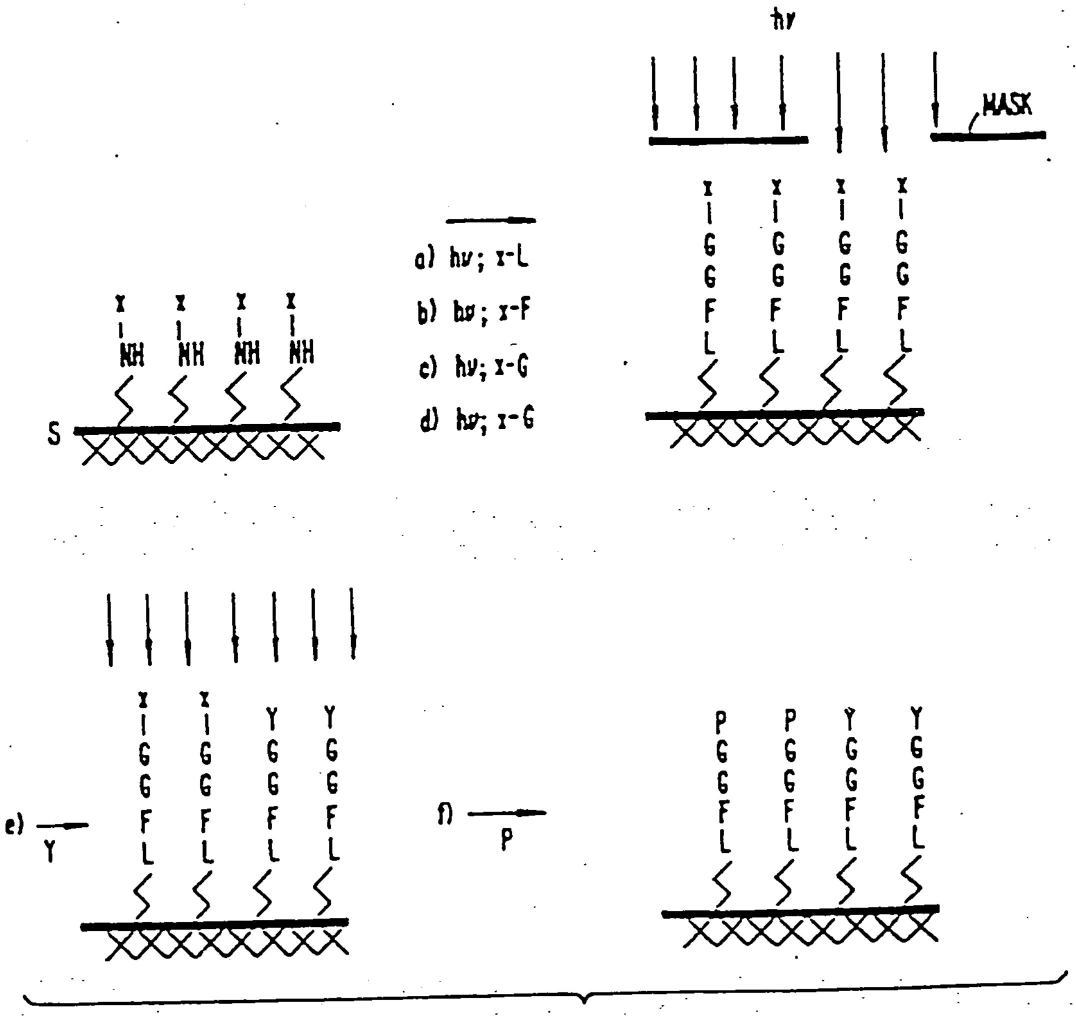


FIG. 2.

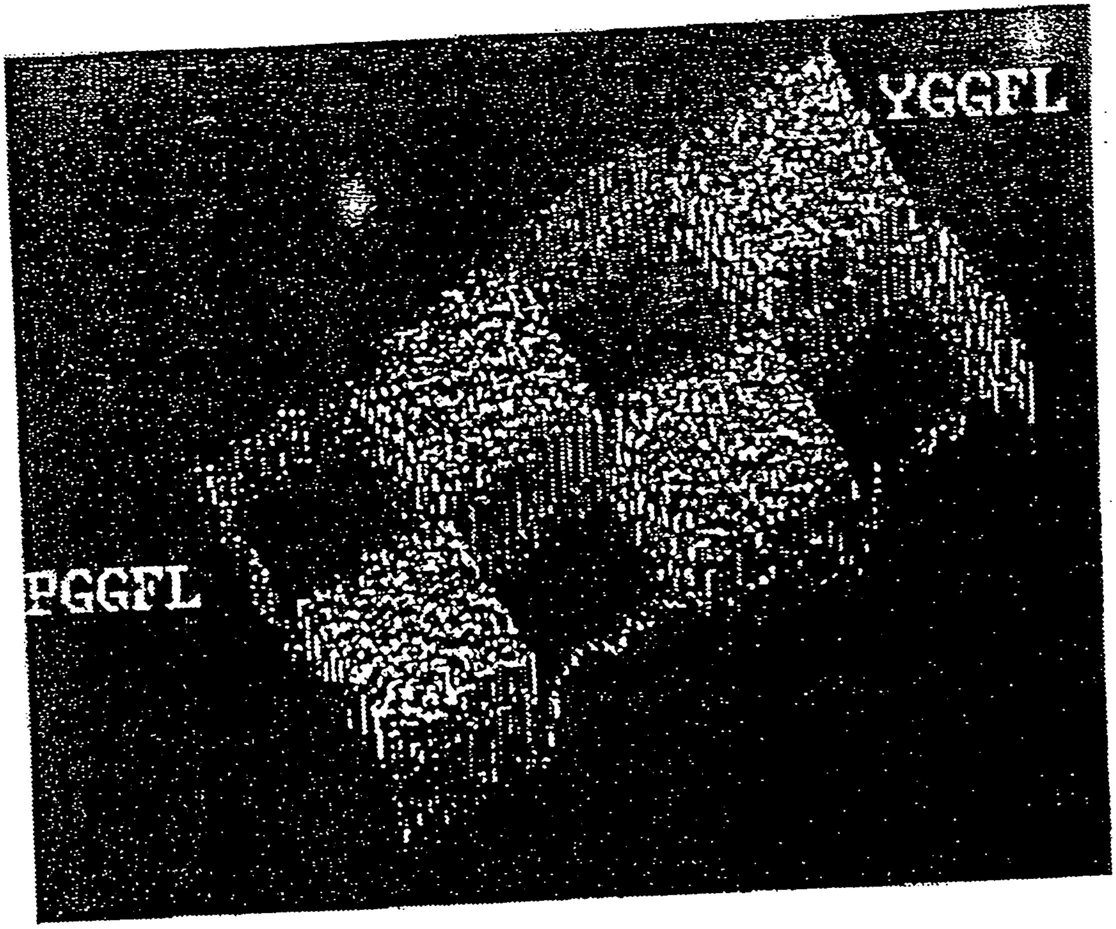


FIG. 3.

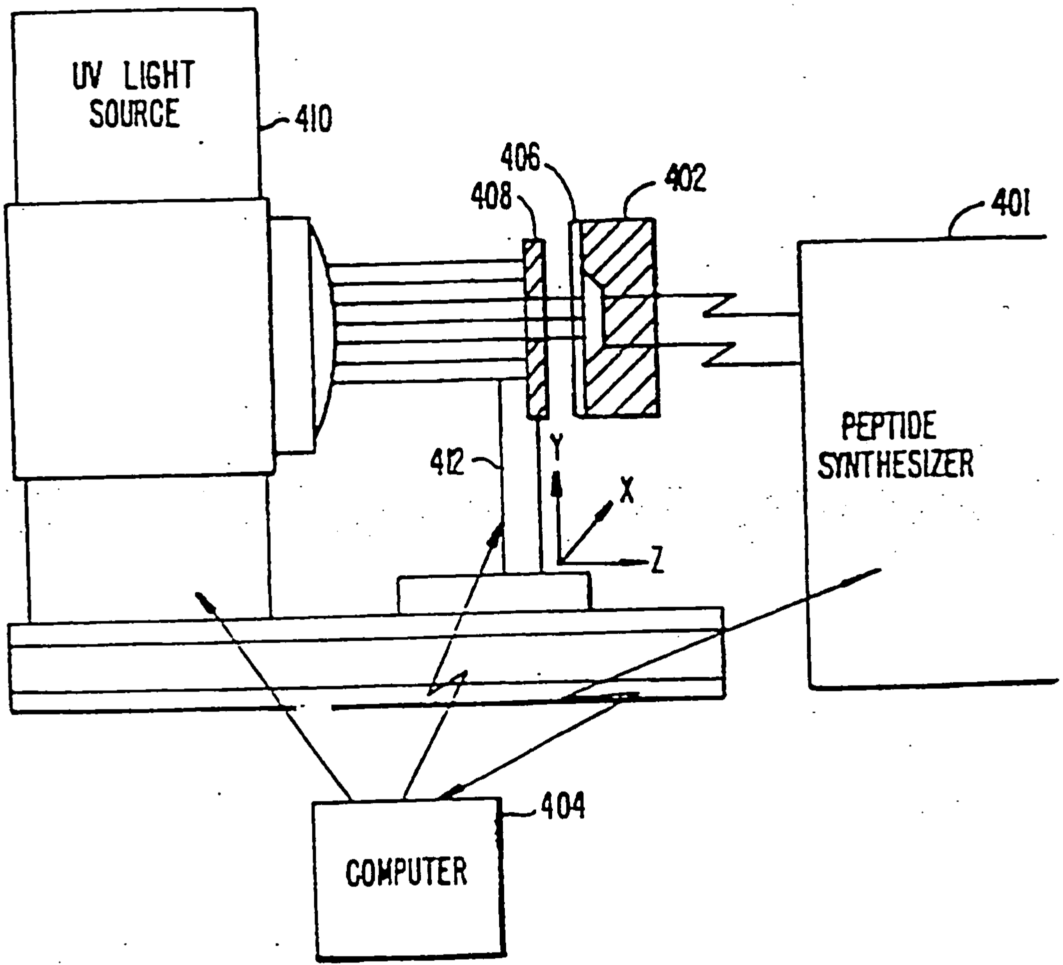

UV LIGHT SOURCE
410
406
408
402
401
412
Y
X
Z
PEPTIDE SYNTHESIZER
404
COMPUTER

FIG. 4.

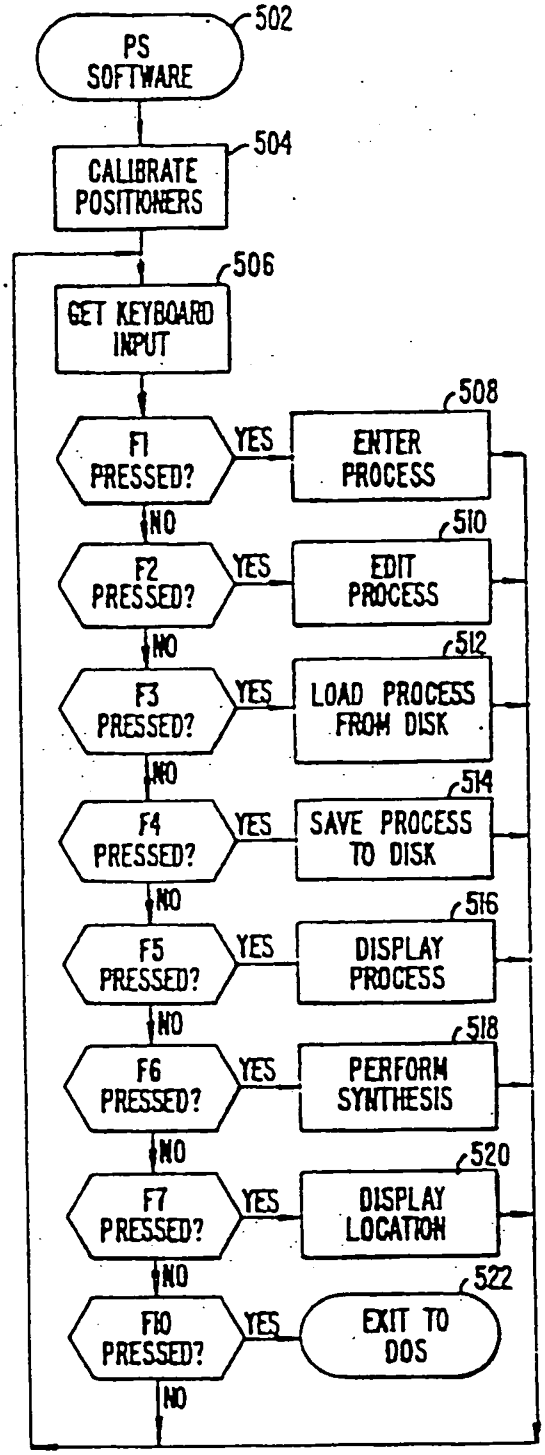

PS SOFTWARE
502
CALIBRATE POSITIONERS
504
GET KEYBOARD INPUT
506
F1 PRESSED?
YES
ENTER PROCESS
508
NO
F2 PRESSED?
YES
EDIT PROCESS
510
NO
F3 PRESSED?
YES
LOAD PROCESS FROM DISK
512
NO
F4 PRESSED?
YES
SAVE PROCESS TO DISK
514
NO
F5 PRESSED?
YES
DISPLAY PROCESS
516
NO
F6 PRESSED?
YES
PERFORM SYNTHESIS
518
NO
F7 PRESSED?
YES
DISPLAY LOCATION
520
NO
F10 PRESSED?
YES
EXIT TO DOS
522
NO

FIG. 5A.

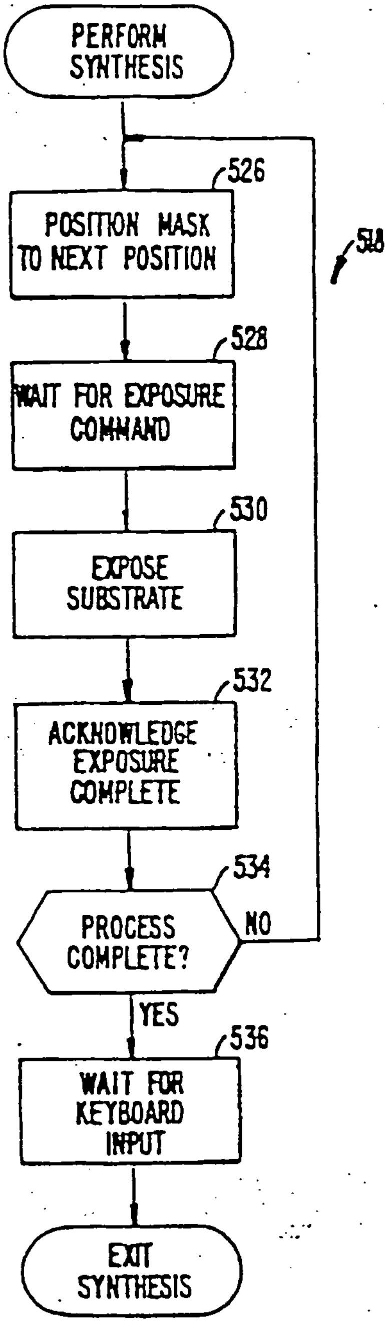

PERFORM SYNTHESIS
POSITION MASK TO NEXT POSITION
526
518
WAIT FOR EXPOSURE COMMAND
528
EXPOSE SUBSTRATE
530
ACKNOWLEDGE EXPOSURE COMPLETE
532
PROCESS COMPLETE?
534
NO
YES
WAIT FOR KEYBOARD INPUT
536
EXIT SYNTHESIS

FIG. 5B.

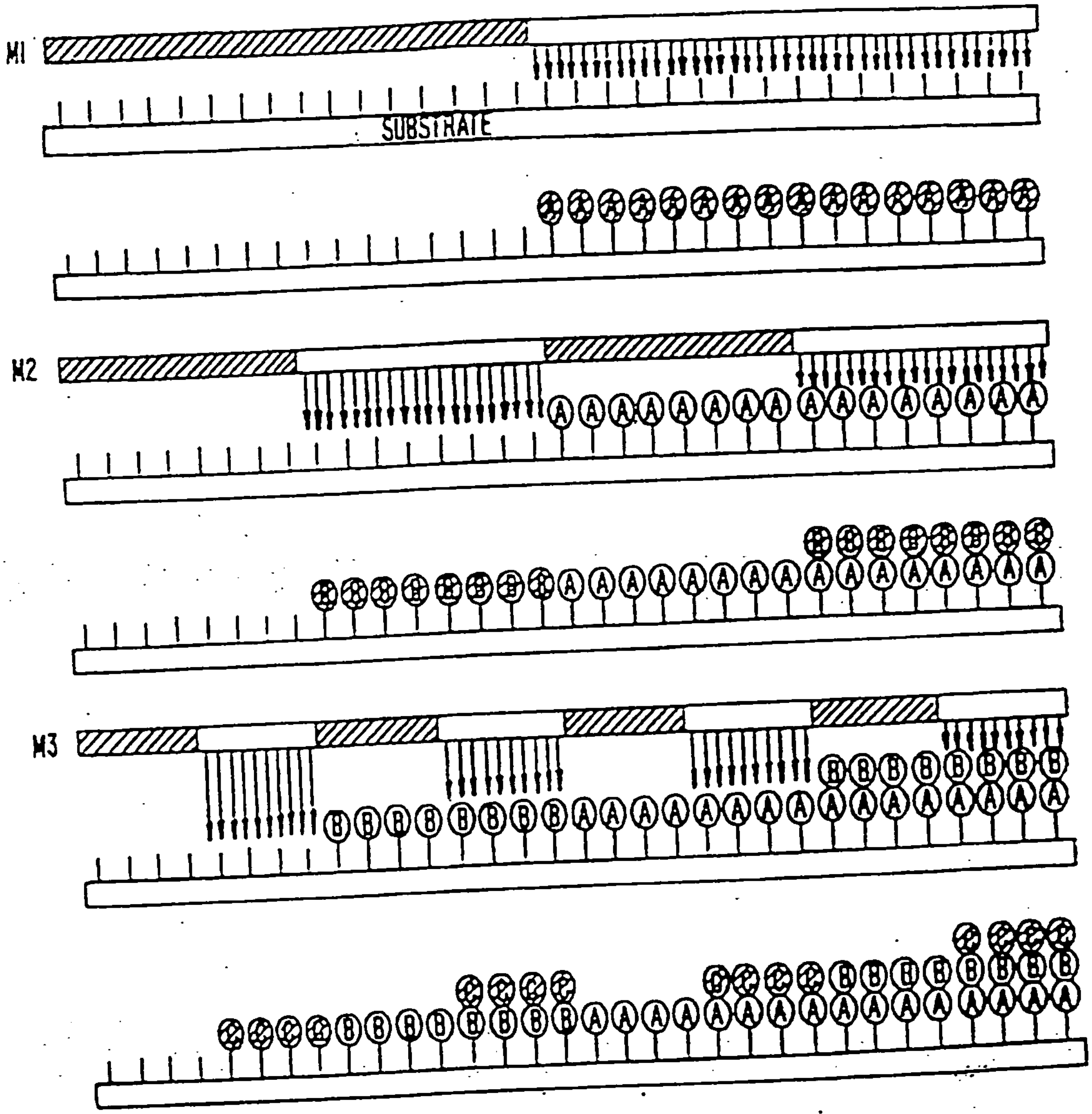


FIG. 6A.

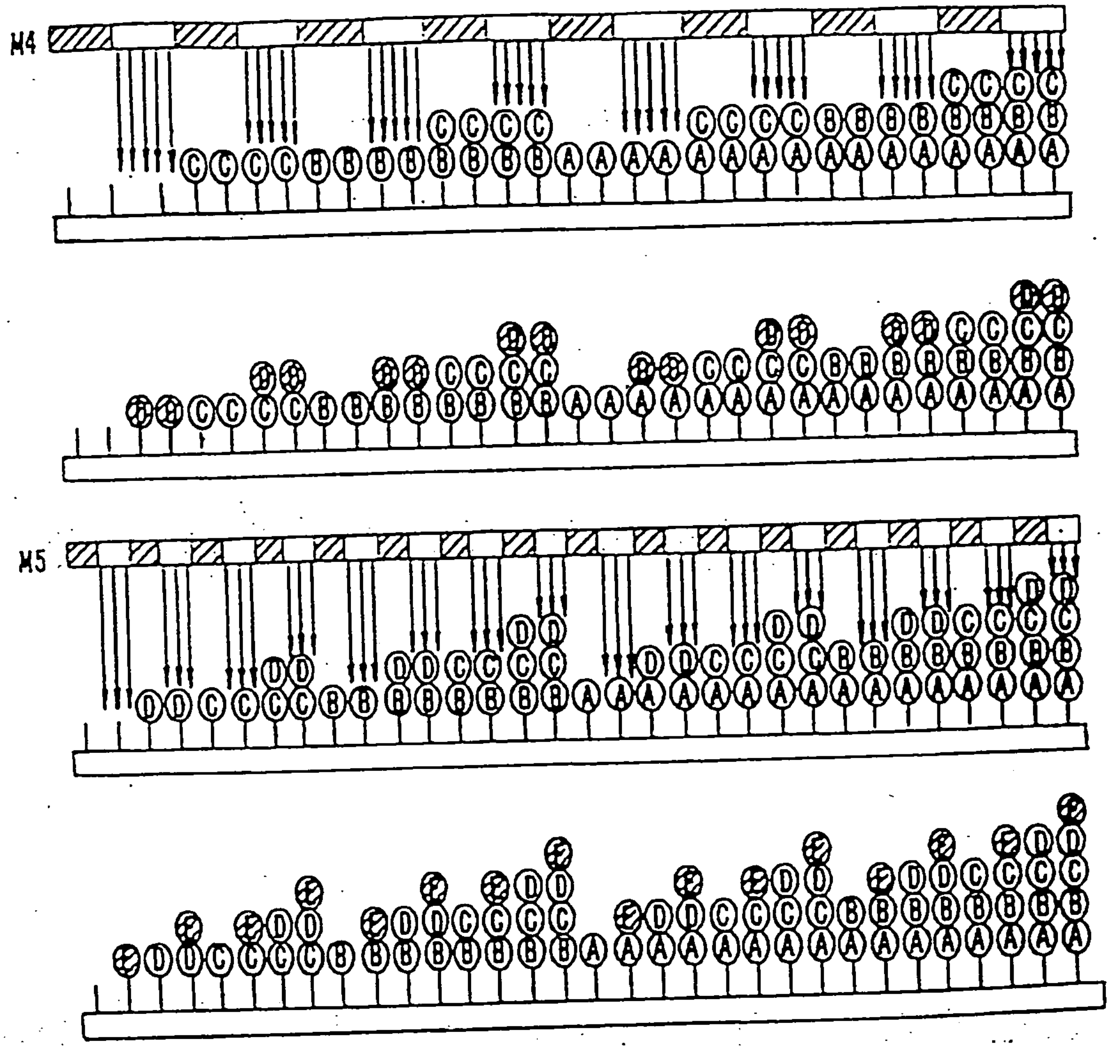


FIG. 6B.

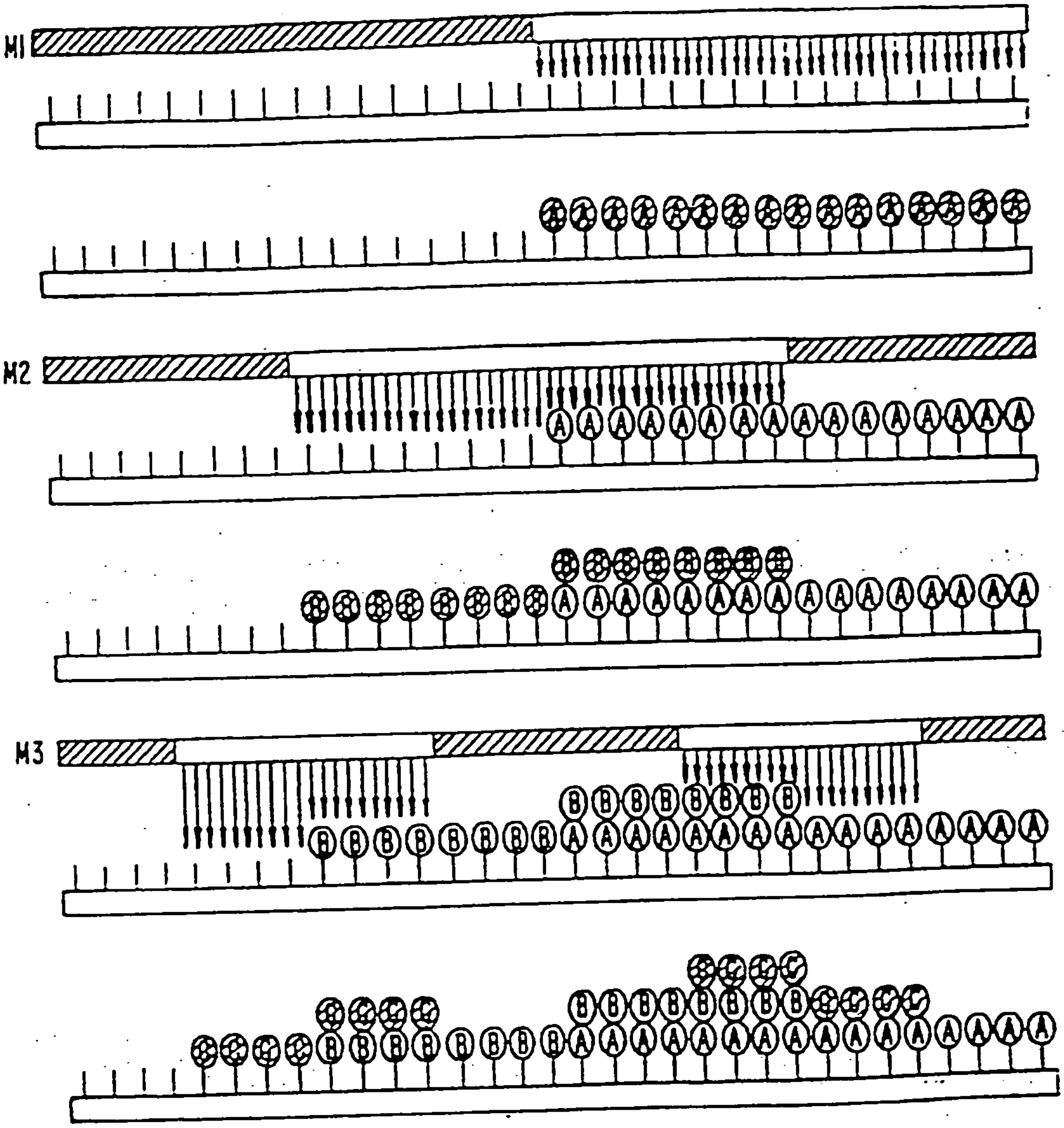


FIG. 7A.

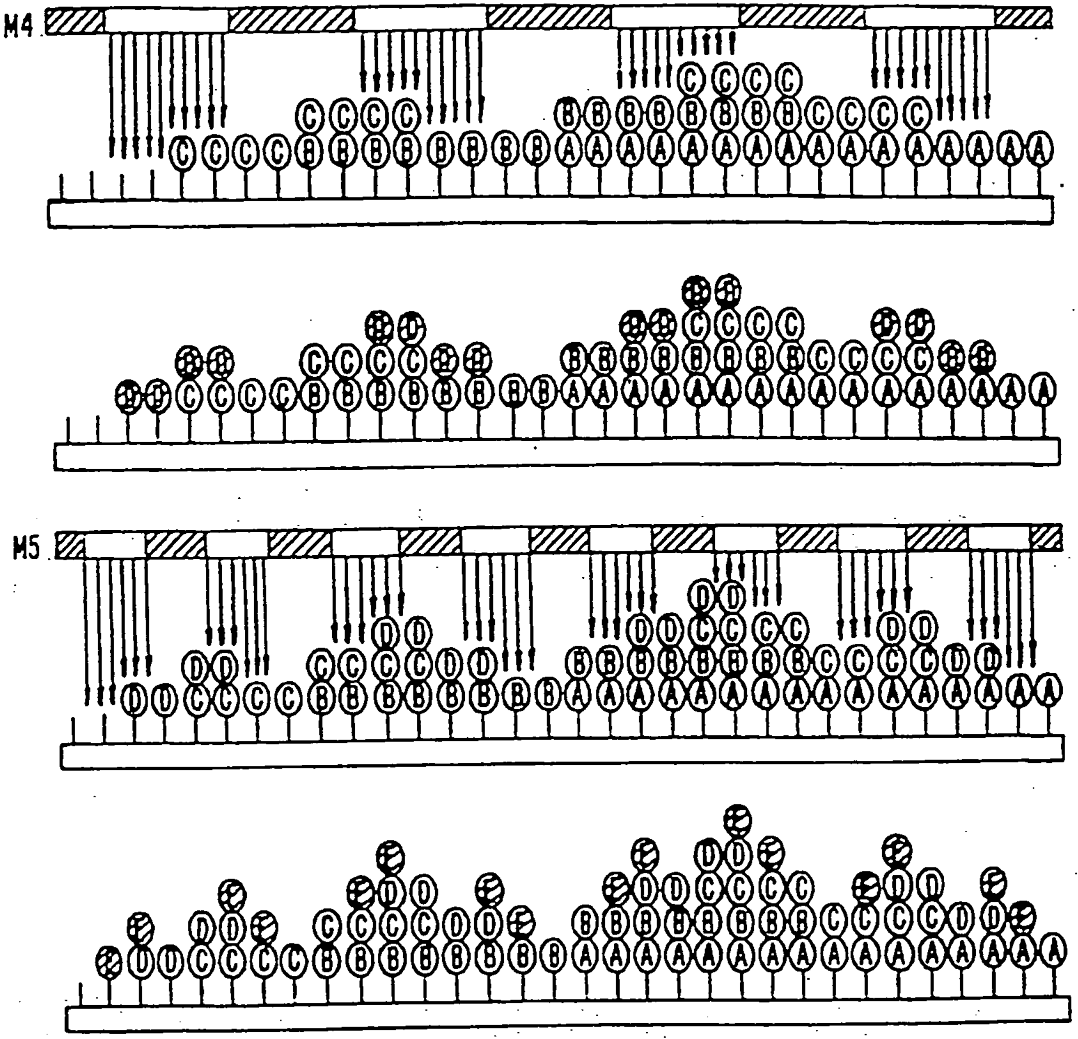


FIG. 7B.

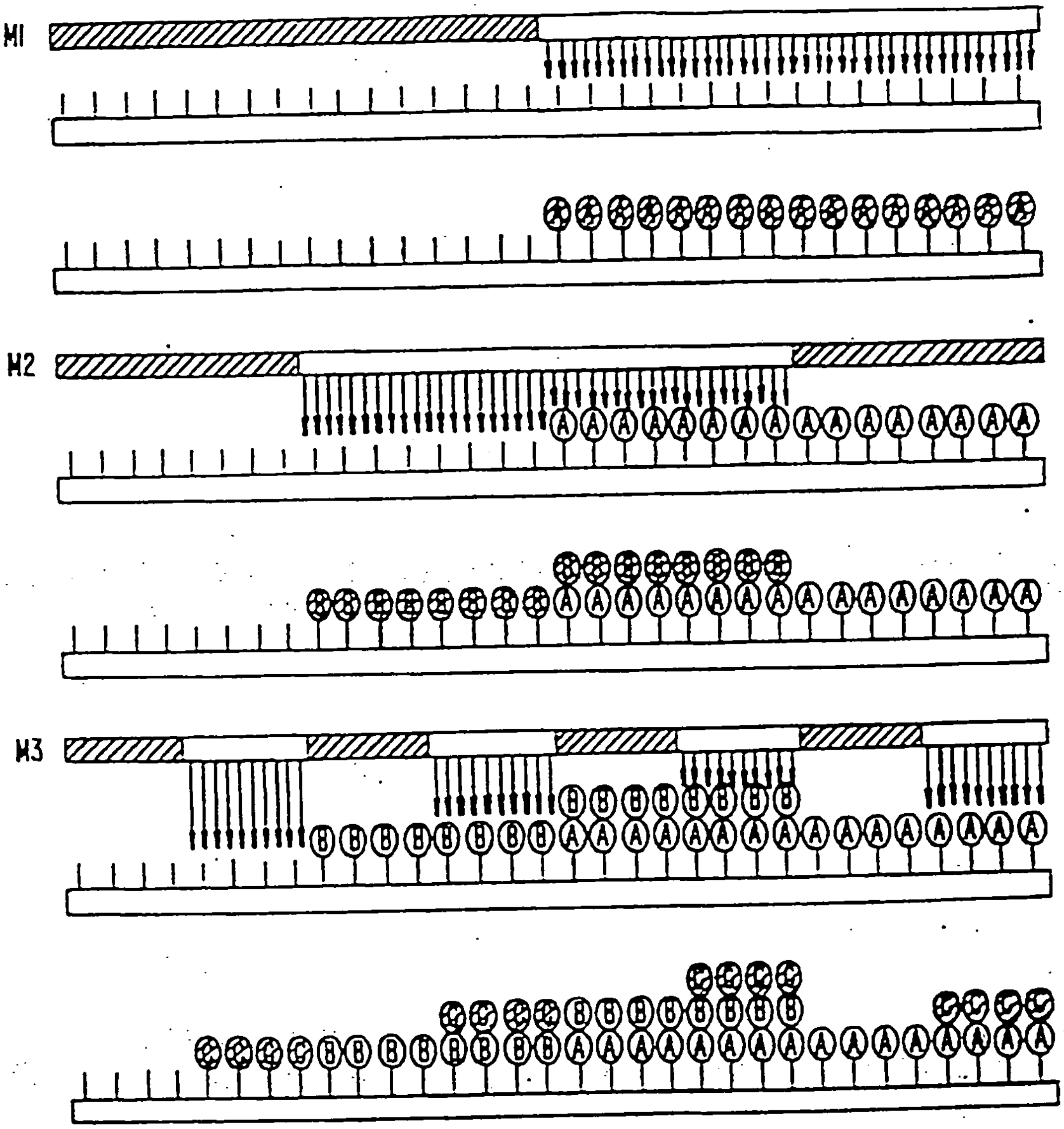


FIG. 8A.

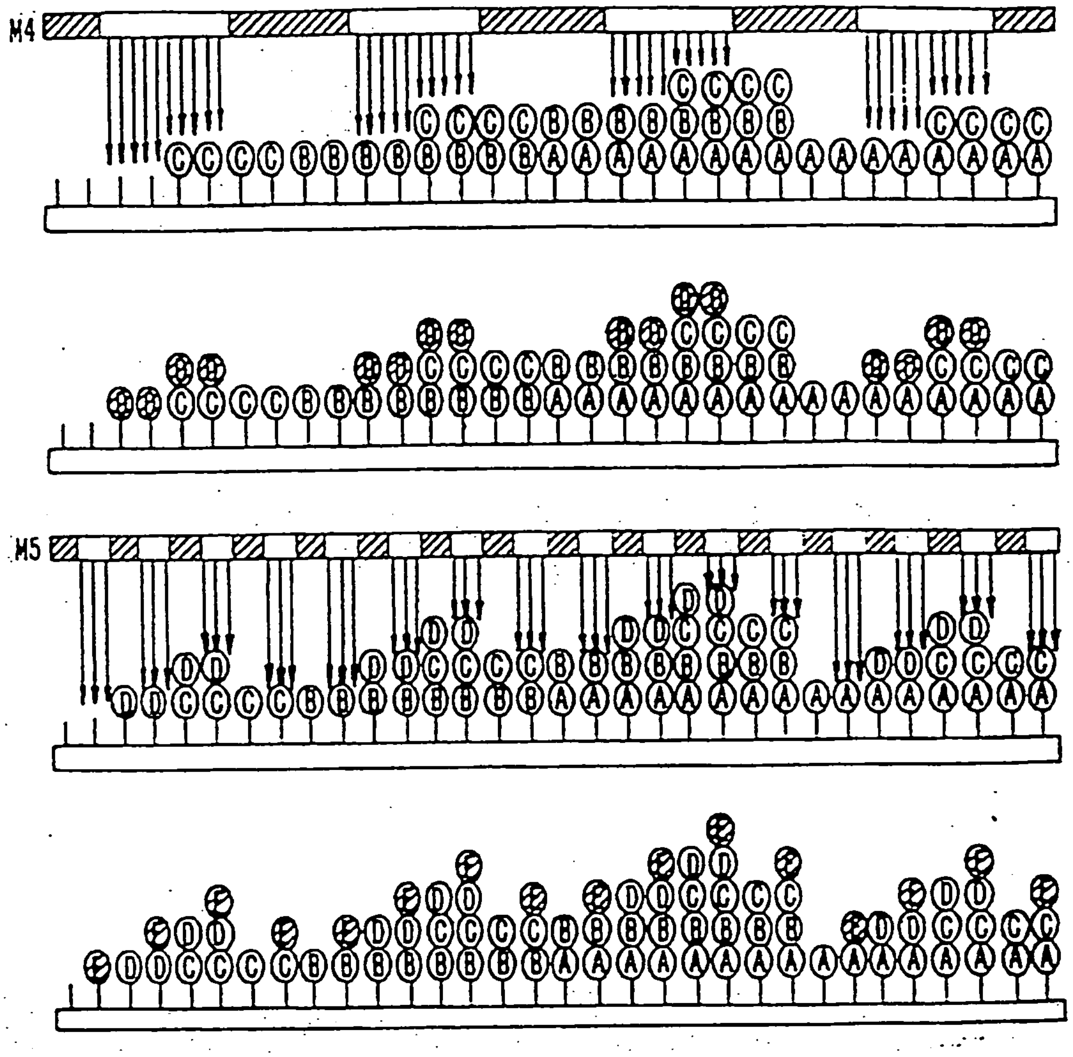


FIG. 8B.

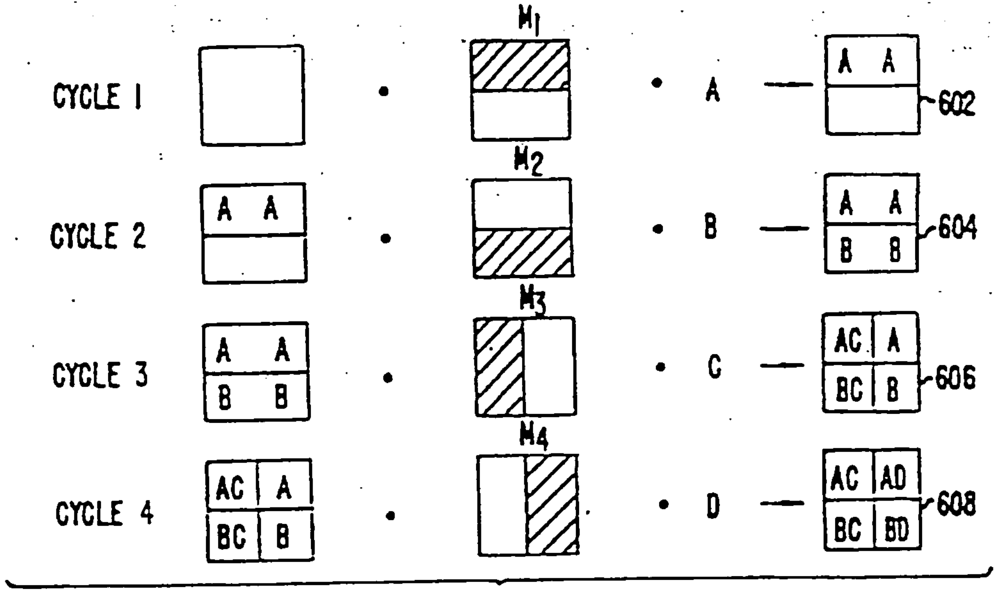

CYCLE 1
M1
A
A A
602
CYCLE 2
A A
M2
B
A A
B B
604
CYCLE 3
A A
B B
M3
C
AC A
BC B
606
CYCLE 4
AC A
BC B
M4
D
AC AD
BC BD
608

FIG. 9A.

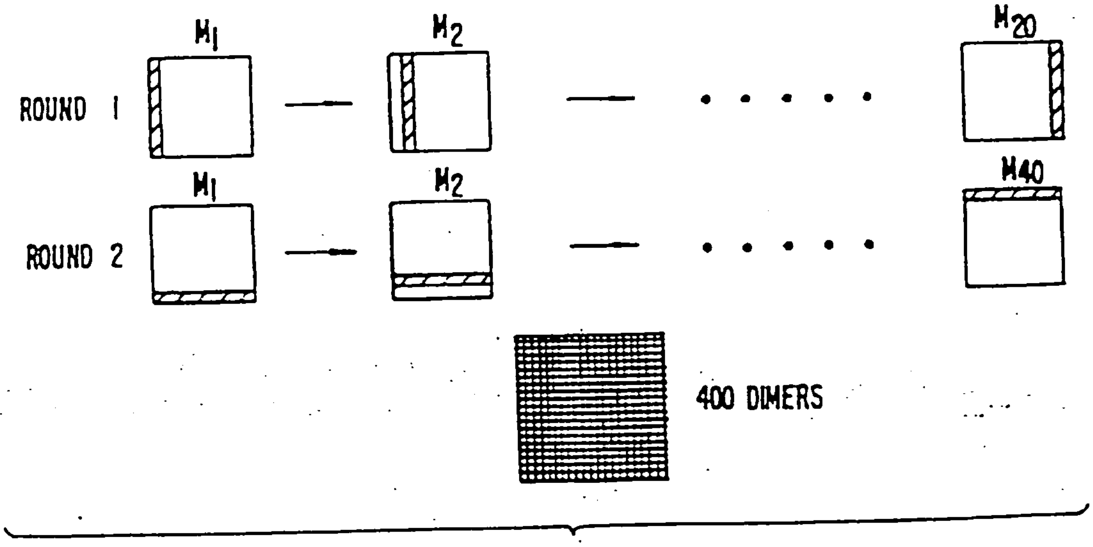

M1
M2
M20
ROUND 1
M1
M2
M40
ROUND 2
400 DIMERS

FIG. 9B.

(x, y)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1		f		f		f		f		f		f		f		f		f		f		f		f		f		f		f		f
2		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y		Y
3			G	G	G	G					G	G	G	G					G	G	G	G					G	G	G	G		
4					A	A	A	A					A	A	A	A					A	A	A	A					A	A	A	A
5			G	G	G	G					G	G	G	G					G	G	G	G					G	G	G	G		
6					T	T	T	T					T	T	T	T					T	T	T	T					T	T	T	T
7									F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F								
8																	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
9									S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S								
10																	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

FIG. 10.

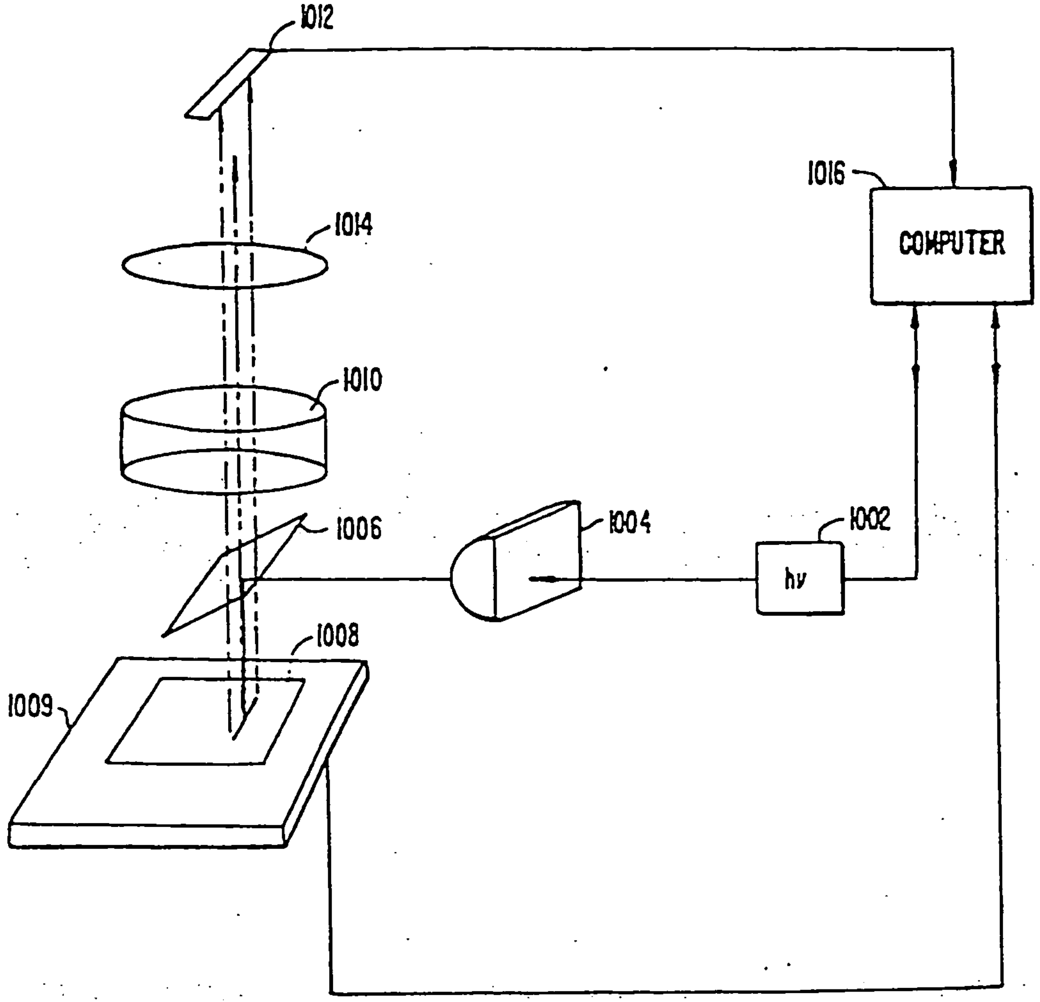


FIG. 11.

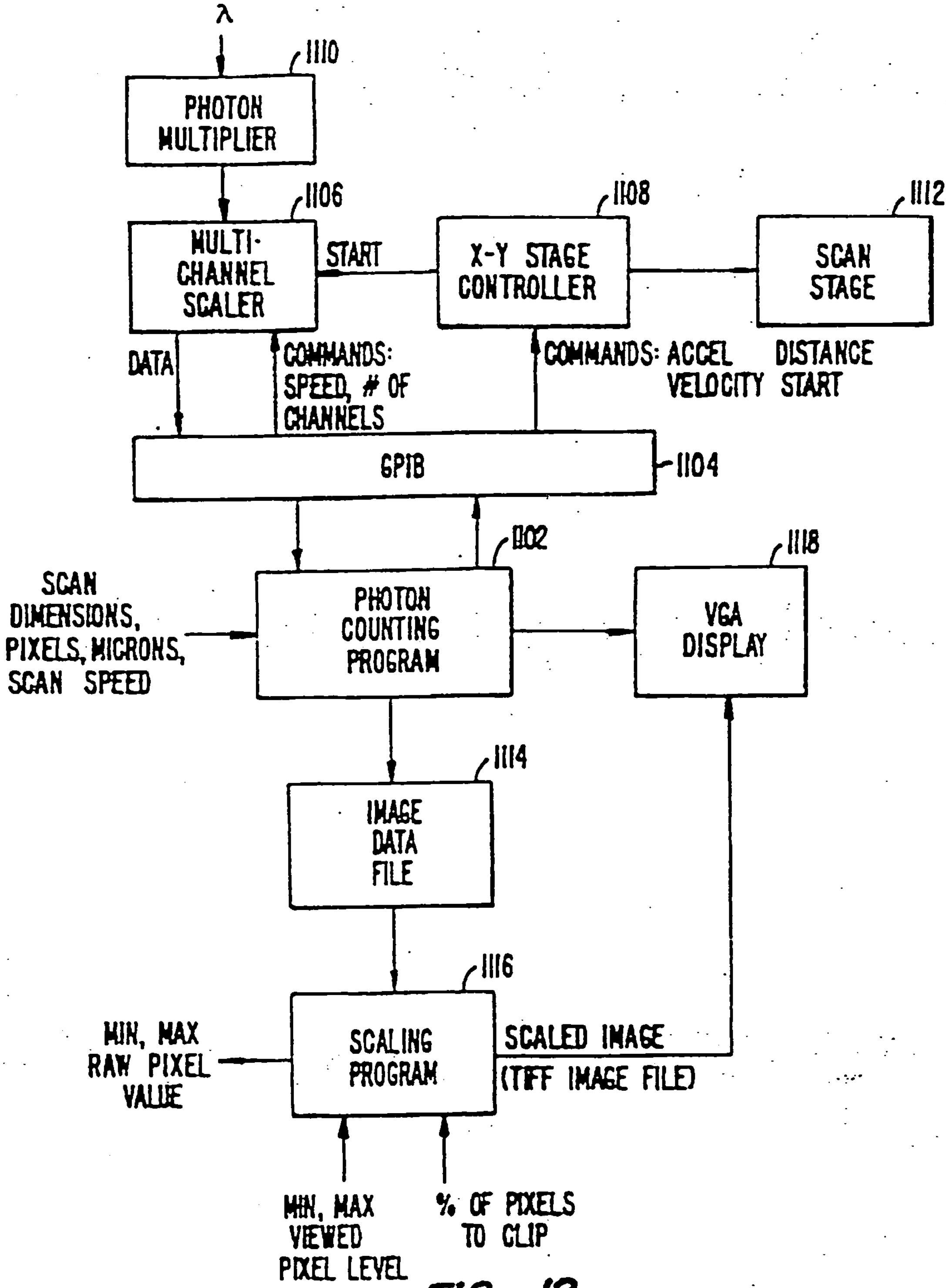

λ
1110
PHOTON MULTIPLIER
1106
MULTI-CHANNEL SCALER
START
1108
X-Y STAGE CONTROLLER
1112
SCAN STAGE
DATA
COMMANDS: SPEED, # OF CHANNELS
COMMANDS: ACCEL VELOCITY DISTANCE START
GPIB
1104
1102
1118
SCAN DIMENSIONS, PIXELS, MICRONS, SCAN SPEED
PHOTON COUNTING PROGRAM
VGA DISPLAY
1114
IMAGE DATA FILE
1116
MIN, MAX RAW PIXEL VALUE
SCALING PROGRAM
SCALED IMAGE (TIFF IMAGE FILE)
MIN, MAX VIEWED PIXEL LEVEL
% OF PIXELS TO CLIP

FIG. 12.

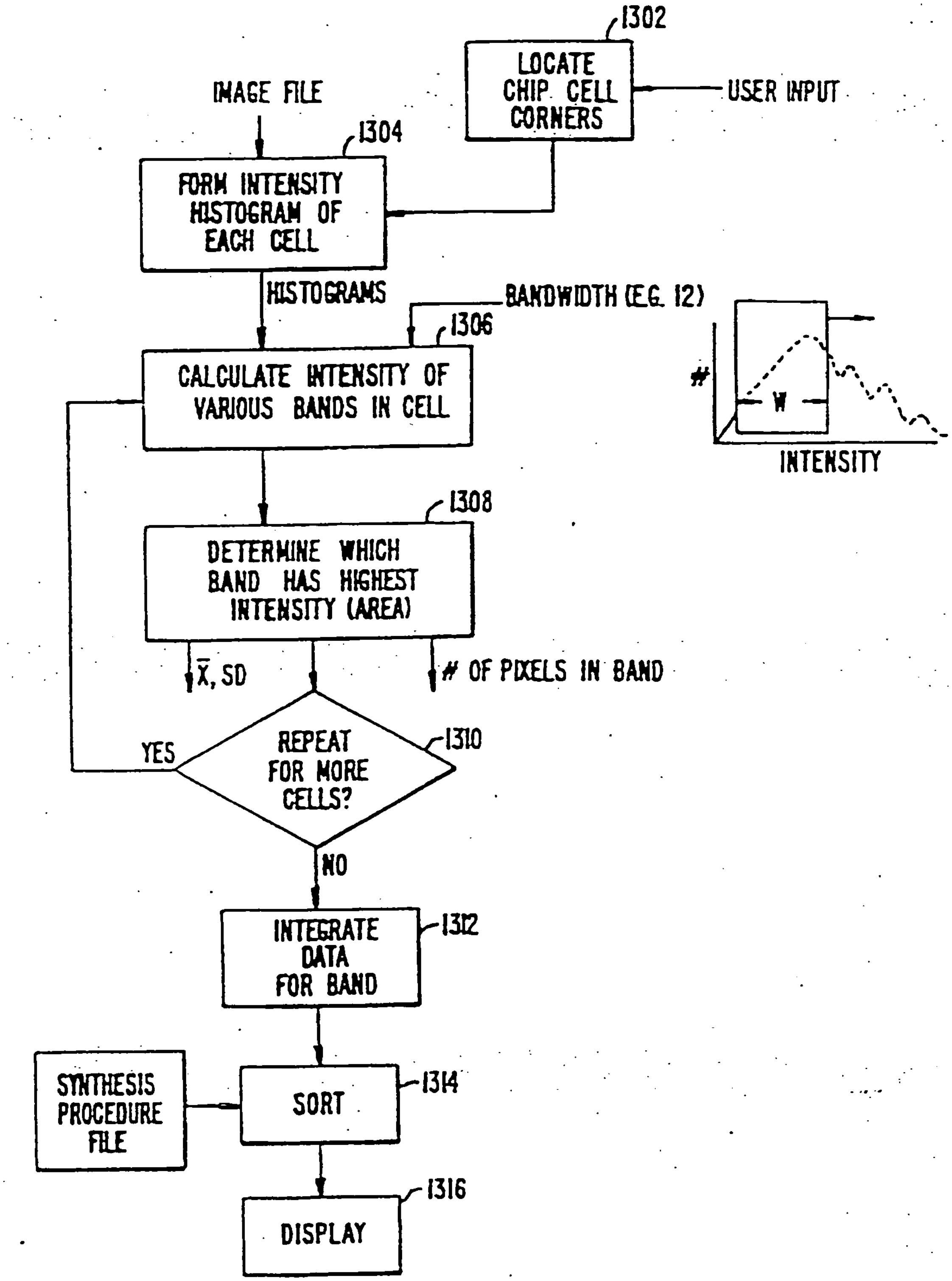

1302
LOCATE CHIP CELL CORNERS
USER INPUT
IMAGE FILE
1304
FORM INTENSITY HISTOGRAM OF EACH CELL
HISTOGRAMS
BANDWIDTH (E.G. 12)
1306
CALCULATE INTENSITY OF VARIOUS BANDS IN CELL
#
W
INTENSITY
1308
DETERMINE WHICH BAND HAS HIGHEST INTENSITY (AREA)
$\overline{X}$, SD
# OF PIXELS IN BAND
YES
REPEAT FOR MORE CELLS?
1310
NO
INTEGRATE DATA FOR BAND
1312
SYNTHESIS PROCEDURE FILE
SORT
1314
1316
DISPLAY

FIG. 13.

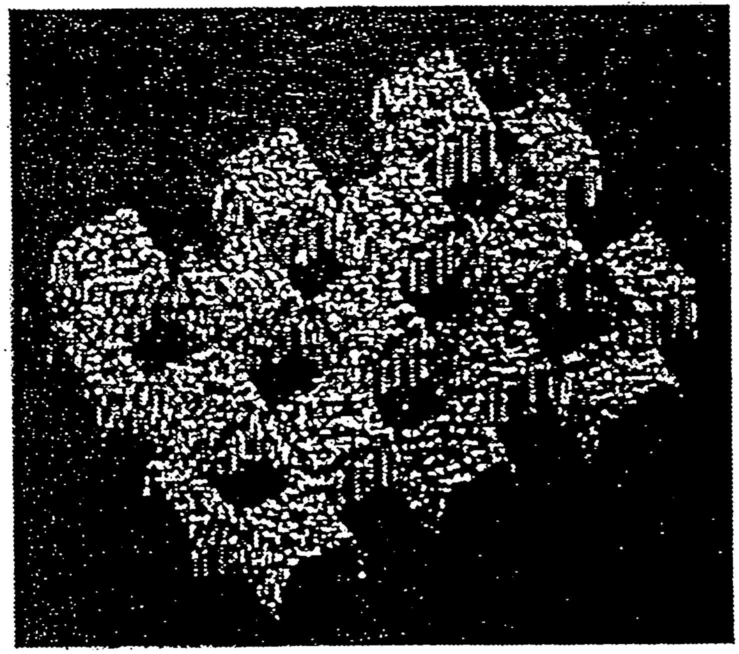


FIG. 14.

ARRAYS OF MATERIALS ATTACHED TO A SUBSTRATE

CROSS REFERENCE TO RELATED APPLICATIONS

This application is a division of U.S. patent application Ser. No. 08/390.272. filed Feb. 16. 1995. now U.S. Pat. No. 5.489.678. which is a continuation of U.S. patent application Ser. No. 07/624.120, filed Dec. 6. 1990. now abandoned. which is a continuation-in-part of U.S. patent application Ser. No. 07/492.462, filed Mar. 7, 1990. now U.S. Pat. No. 5.143.854, which is a continuation-in-part of U.S. patent application Ser. No. 07/362.901, filed Jun. 7, 1989. now abandoned. and hereby incorporated herein by reference for all purposes. This application is also a continuation-in-part of U.S. patent application Ser. No. 08/456.887, filed Jun. 1, 1995, which is a division of U.S. patent application Ser. No. 07/954.646. filed Sep. 30, 1992. now U.S. Pat. No. 5.445, 934, which is a division of U.S. patent application Ser. No. 07/850.356. filed Mar. 12, 1992, now U.S. Pat. No. 5.405, 783. which is a division of U.S. patent application Ser. No. 07/492.462, filed Mar. 7. 1990, now U.S. Pat. Ser. No. 5.143.854. which is a continuation-in-part of U.S. patent application Ser. No. 07/362.901 filed Jun. 7, 1989. now abandoned.

This application is also related to U.S. patent application Ser. No. 08/670.118 filed Jun. 25, 1996, which is a division of U.S. patent application Ser. No. 08/168,104, filed Dec. 15, 1993, which is a continuation of U.S. patent application Ser. No. 07/624.114. filed Dec. 6, 1990, now abandoned, and U.S. patent application Ser. No. 07/626.730, filed Dec. 6. 1990, now U.S. Pat. No. 5.547.839, and also incorporated herein by reference for all purposes.

COPYRIGHT NOTICE

BACKGROUND OF THE INVENTION

The present invention relates to the field of polymer synthesis. More specifically, the invention provides a reactor system, a masking strategy, photoremovable protective groups, data collection and processing techniques, and applications for light directed synthesis of diverse polymer sequences on substrates.

SUMMARY OF THE INVENTION

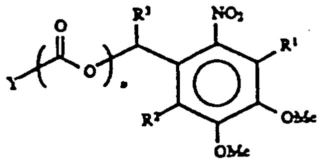
Methods, apparatus, and compositions for synthesis and use of diverse polymer sequences on a substrate are disclosed, as well as applications thereof.

According to one aspect of the invention, an improved reactor system for synthesis of diverse polymer sequences on a substrate is provided. According to this embodiment the invention provides for a reactor for contacting reaction fluids to a substrate; a system for delivering selected reaction fluids to the reactor; a translation stage for moving a mask or substrate from at least a first relative location relative to a second relative location; a light for illuminating the substrate through a mask at selected times; and an appropriately programmed digital computer for selectively directing a

flow of fluids from the reactor system. selectively activating the translation stage, and selectively illuminating the substrate so as to form a plurality of diverse polymer sequences on the substrate at predetermined locations.

The invention also provides a technique for selection of linker molecules in a very large scale immobilized polymer synthesis (VLSIPS™) method. According to this aspect of the invention, the invention provides a method of screening a plurality of linker polymers for use in binding affinity studies. The invention includes the steps of forming a plurality of linker polymers on a substrate in selected regions, the linker polymers formed by the steps of recursively: on a surface of a substrate, irradiating a portion of the selected regions to remove a protective group, and contacting the surface with a monomer; contacting the plurality of linker polymers with a ligand; and contacting the ligand with a labeled receptor.

According to another aspect of the invention, improved photoremovable protective groups are provided. According to this aspect of the invention a compound having the formula:

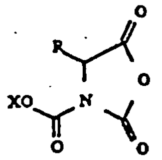


wherein n=0 or 1; Y is selected from the group consisting of an oxygen of the carboxyl group of a natural or unnatural amino acid, an amino group of a natural or unnatural amino acid, or the C-5' oxygen group of a natural or unnatural deoxyribonucleic or ribonucleic acid; R^1 and R^2 independently are a hydrogen atom, a lower alkyl, aryl, benzyl, halogen, hydroxyl, alkoxyl, thiol, thioether, amino, nitro, carboxyl, formate, formamido, sulfido, or phosphido group; and R^3 is a alkoxy, alkyl, aryl, hydrogen, or alkenyl group is provided.

The invention also provides improved masking techniques for the VLSIPS™ methodology. According to one aspect of the masking technique, the invention provides an ordered method for forming a plurality of polymer sequences by sequential addition of reagents comprising the step of serially protecting and deprotecting portions of the plurality of polymer sequences for addition of other portions of the polymer sequences using a binary synthesis strategy.

Improved data collection equipment and techniques are also provided. According to one embodiment, the instrumentation provides a system for determining affinity of a receptor to a ligand comprising: means for applying light to a surface of a substrate, the substrate comprising a plurality of ligands at predetermined locations, the means for providing simultaneous illumination at a plurality of the predetermined locations; and an array of detectors for detecting light fluoresced at the plurality of predetermined locations. The invention further provides for improved data analysis techniques including the steps of exposing fluorescently labelled receptors to a substrate, the substrate comprising a plurality of ligands in regions at known locations; at a plurality of data collection points within each of the regions, determining an amount of light fluoresced from the data collection points; removing the data collection points deviating from a predetermined statistical distribution; and determining a relative binding affinity of the receptor to remaining data collection points.

Protected amino acid N-carboxy anhydrides for use in polymer synthesis are also disclosed. According to this aspect, the invention provides a compound having the formula:



where R is a side chain of a natural or unnatural amino acid and X is a photoremovable protecting group.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 schematically illustrates light-directed spatially-addressable parallel chemical synthesis;

FIG. 2 schematically illustrates one example of light-directed peptide synthesis;

FIG. 3 is a three-dimensional representation of a portion of the checkerboard array of YGGFL and[PGGFL;

FIG. 4 schematically illustrates an automated system for synthesizing diverse polymer sequences;

FIGS. 5*a* and 5*b* illustrate operation of a program for polymer sythesis;

FIGS. 6*a* and 6*b* are a schematic illustration of a "pure" binary masking strategy;

FIGS. 7*a* and 7*b* are a schematic illustration of a gray code binary masking strategy;

FIGS. 8*a* and 8*b* are a schematic illustration of a modified gray code binary masking strategy;

FIG. 9*a* schematically illustrates a masking scheme for a four step synthesis;

FIG. 9*b* schematically illustrates synthesis of all 400 peptide dimers;

FIG. 10 is a coordinate map for the ten-step binary synthesis;

FIG. 11 schematically illustrates a data collection system;

FIG. 12 is a block diagram illustrating the architecture of the data collection system;

FIG. 13 is a flow chart illustrating operation of software for the data collection/analysis system; and

FIG. 14 illustrates a three-dimensional plot of intensity versus position for light directed synthesis of a dinucleotide.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

CONTENTS

I. DEFINITIONS

Certain terms used herein are intended to have the following general definitions:

1. Complementary:

Refers to the topological compatibility or matching together of interacting surfaces of a ligand molecule and its receptor. Thus, the receptor and its ligand can be described as complementary, and furthermore, the contact surface characteristics are complementary to each other.

2. Epitope:

The portion of an antigen molecule which is delineated by the area of interaction with the subclass of receptors known as antibodies.

3. Ligand:

A ligand is a molecule that is recognized by a particular receptor. Examples of ligands that can be investigated by this invention include, but are not restricted to, agonists and antagonists for cell membrane receptors, toxins and venoms, viral epitopes, hormones, hormone receptors, peptides, enzymes, enzyme substrates, cofactors, drugs (e.g. opiates, steriods, etc.), lectins, sugars, oligonucleotides, nucleic acids, oligosaccharides, proteins, and monoclonal antibodies.

4. Monomer:

A member of the set of small molecules which can be joined together to form a polymer. The set of monomers includes but is not restricted to, for example, the set of common L-amino acids, the set of D-amino acids, the set of synthetic amino acids, the set of nucleotides and the set of pentoses and hexoses. As used herein, monomers refers to any member of a basis set for synthesis of a polymer. For example, dimers of the 20 naturally occurring L-amino acids form a basis set of 400 monomers for synthesis of polypeptides. Different basis sets of monomers may be used at successive steps in the synthesis of a polymer. Furthermore, each of the sets may include protected members which are modified after synthesis.

5. Peptide:

A polymer in which the monomers are alpha amino acids and which are joined together through amide bonds and alternatively referred to as a polypeptide. In the context of this specification it should be appreciated that the amino acids may be the L-optical isomer or the D-optical isomer.

Peptides are often two or more amino acid monomers long. and often more than 20 amino acid monomers long. Standard abbreviations for amino acids are used (e.g., P for proline). These abbreviations are included in Stryer. *Biochemistry*. Third Ed., 1988. which is incorporated herein by reference for all purposes.

6. Radiation:

Energy which may be selectively applied including energy having a wavelength of between 10^{-14} and 10^{4} meters including, for example, electron beam radiation. gamma radiation. x-ray radiation. ultraviolet radiation, visible light. infrared radiation. microwave radiation. and radio waves. "Irradiation" refers to the application of radiation to a surface.

7. Receptor:

A molecule that has an affinity for a given ligand. Receptors may be naturally-occurring or manmade molecules. Also, they can be employed in their unaltered state or as aggregates with other species. Receptors may be attached, covalently or noncovalently, to a binding member, either directly or via a specific binding substance. Examples of receptors which can be employed by this invention include, but are not restricted to, antibodies, cell membrane receptors, monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells or other materials), drugs, polynucleotides, nucleic acids, peptides, cofactors, lectins, sugars, polysaccharides, cells, cellular membranes, and organelles. Receptors are sometimes referred to in the art as anti-ligands. As the term receptors is used herein, no difference in meaning is intended. A "Ligand Receptor Pair" is formed when two macromolecules have combined through molecular recognition to form a complex. Other examples of receptors which can be investigated by this invention include but are not restricted to:

a) Microorqanism receptors:

Determination of ligands which bind to receptors, such as specific transport proteins or enzymes essential to survival of microorganisms, is useful in developing a new class of antibiotics. Of particular value would be antibiotics against opportunistic fungi, protozoa, and those bacteria resistant to the antibiotics in current use.

b) Enzymes:

For instance, one type of receptor is the binding site of enzymes such as the enzymes responsible for cleaving neurotransmitters; determination of ligands which bind to certain receptors to modulate the action of the enzymes which cleave the different neurotransmitters is useful in the development of drugs which can be used in the treatment of disorders of neurotransmission.

c) Antibodies:

For instance, the invention may be useful in investigating the ligand-binding site on the antibody molecule which combines with the epitope of an antigen of interest; determining a sequence that mimics an antigenic epitope may lead to the development of vaccines of which the immunogen is based on one or more of such sequences or lead to the development of related diagnostic agents or compounds useful in therapeutic treatments such as for auto-immune diseases (e.g., by blocking the binding of the "self" antibodies).

d) Nucleic Acids:

Sequences of nucleic acids may be synthesized to establish DNA or RNA binding sequences.

e) Catalytic Polypeptides:

Polymers, preferably polypeptides, which are capable of promoting a chemical reaction involving the conversion of one or more reactants to one or more products. Such polypeptides generally include a binding site specific for at least one reactant or reaction intermediate and an active functionality proximate to the binding site, which functionality is capable of chemically modifying the bound reactant. Catalytic polypeptides are described in, for example, U.S. Pat. No. 5.215.899, which is incorporated herein by reference for all purposes.

f) Hormone receptors:

Examples of hormones receptors include, e.g., the receptors for insulin and growth hormone. Determination of the ligands which bind with high affinity to a receptor is useful in the development of, for example, an oral replacement of the daily injections which diabetics must take to relieve the symptoms of diabetes, and in the other case, a replacement for the scarce human growth hormone which can only be obtained from cadavers or by recombinant DNA technology. Other examples are the vasoconstrictive hormone receptors; determination of those ligands which bind to a receptor may lead to the development of drugs to control blood pressure.

g) Opiate receptors:

Determination of ligands which bind to the opiate receptors in the brain is useful in the development of less-addictive replacements for morphine and related drugs.

8. Substrate:

A material having a rigid or semi-rigid surface. In many embodiments, at least one surface of the substrate will be substantially flat, although in some embodiments it may be desirable to physically separate synthesis regions for different polymers with, for example, wells, raised regions, etched trenches, or the like. According to other embodiments, small beads may be provided on the surface which may be released upon completion of the synthesis.

9. Protective Group:

A material which is chemically bound to a monomer unit and which may be removed upon selective exposure to an activator such as electromagnetic radiation. Examples of protective groups with utility herein include those comprising nitropiperonyl, pyrenylmethoxy-carbonyl, nitroveratryl, nitrobenzyl, dimethyl dimethoxybenzyl, 5-bromo-7-nitroindolinyl, o-hydroxy-α-methyl cinnamoyl, and 2-oxymethylene anthraquinone.

10. Predefined Region:

A predefined region is a localized area on a surface which is, was, or is intended to be activated for formation of a polymer. The predefined region may have any convenient shape, e.g., circular, rectangular, elliptical, wedge-shaped, etc. For the sake of brevity herein, "predefined regions" are sometimes referred to simply as "regions."

11. Substantially Pure:

A polymer is considered to be "substantially pure" within a predefined region of a substrate when it exhibits characteristics that distinguish it from other predefined regions. Typically, purity will be measured in terms of biological activity or function as a result of uniform sequence. Such characteristics will typically be measured by way of binding with a selected ligand or receptor.

12. Activator refers to an energy source adapted to render a group active and which is directed from a source to a predefined location on a substrate. A primary illustration of

an activator is light. Other examples of activators include ion beams, electric fields, magnetic fields, electron beams, x-ray, and the like.

13. Binary Synthesis Strategy refers to an ordered strategy for parallel synthesis of diverse polymer sequences by sequential addition of reagents which may be represented by a reactant matrix, and a switch matrix, the product of which is a product matrix. A reactant matrix is a 1×n matrix of the building blocks to be added. The elements of the switch matrix are binary numbers. In preferred embodiments, a binary strategy is one in which at least two successive steps illuminate half of a region of interest on the substrate. In most preferred embodiments, binary synthesis refers to a synthesis strategy which also factors a previous addition step. For example, a strategy in which a switch matrix for a masking strategy halves regions that were previously illuminated, illuminating about half of the previously illuminated region and protecting the remaining half (while also protecting about half of previously protected regions and illuminating about half of previously protected regions). It will be recognized that binary rounds may be interspersed with non-binary rounds and that only a portion of a substrate may be subjected to a binary scheme, but will still be considered to be a binary masking scheme within the definition herein. A binary "masking" strategy is a binary synthesis which uses light to remove protective groups from materials for addition of other materials such as amino acids. In preferred embodiments, selected columns of the switch matrix are arranged in order of increasing binary numbers in the columns of the switch matrix.

14. Linker refers to a molecule or group of molecules attached to a substrate and spacing a synthesized polymer from the substrate for exposure/binding to a receptor.

II. General

The present invention provides synthetic strategies and devices for the creation of large scale chemical diversity. Solid-phase chemistry, photolabile protecting groups, and photolithography are brought together to achieve light-directed spatially-addressable parallel chemical synthesis in preferred embodiments.

The invention is described herein for purposes of illustration primarily with regard to the preparation of peptides and nucleotides, but could readily be applied in the preparation of other polymers. Such polymers include, for example, both linear and cyclic polymers of nucleic acids, polysaccharides, phospholipids, and peptides having either α-, β-, or ω-amino acids, heteropolymers in which a known drug is covalently bound to any of the above, polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneimines, polyarylene sulfides, polysiloxanes, polyimides, polyacetates, or other polymers which will be apparent upon review of this disclosure. It will be recognized further, that illustrations herein are primarily with reference to C- to N-terminal synthesis, but the invention could readily be applied to N- to C-terminal synthesis without departing from the scope of the invention.

A. Deprotection and Addition

The present invention uses a masked light source or other activator to direct the simultaneous synthesis of many different chemical compounds. FIG. 1 is a flow chart illustrating the process of forming chemical compounds according to one embodiment of the invention. Synthesis occurs on a solid support 2. A pattern of illumination through a mask 4*a* using a light source 6 determines which regions of the support are activated for chemical coupling. In one preferred embodiment activation is accomplished by using light to remove photolabile protecting groups from selected areas of the substrate.

After deprotection, a first of a set of building blocks (indicated by "A" in FIG. 1), each bearing a photolabile protecting group (indicated by "X") is exposed to the surface of the substrate and it reacts with regions that were addressed by light in the preceding step. The substrate is then illuminated through a second mask 4*b*, which activates another region for reaction with a second protected building block "B". The pattern of masks used in these illuminations and the sequence of reactants define the ultimate products and their locations, resulting in diverse sequences at predefined locations, as shown with the sequences ACEG and BDFH in the lower portion of FIG. 1. Preferred embodiments of the invention take advantage of combinatorial masking strategies to form a large number of compounds in a small number of chemical steps.

A high degree of miniaturization is possible because the density of compounds is determined largely with regard to spatial addressability of the activator, in one case the diffraction of light. Each compound is physically accessible and its position is precisely known. Hence, the array is spatially-addressable and its interactions with other molecules can be assessed.

In a particular embodiment shown in FIG. 1, the substrate contains amino groups that are blocked with a photolabile protecting group. Amino acid sequences are made accessible for coupling to a receptor by removal of the photoprotective groups.

When a polymer sequence to be synthesized is, for example, a polypeptide, amino groups at the ends of linkers attached to a glass substrate are derivatized with nitroveratryloxycarbonyl (NVOC), a photoremovable protecting group. The linker molecules may be, for example, aryl acetylene, ethylene glycol oligomers containing from 2–10 monomers, diamines, diacids, amino acids, or combinations thereof. Photodeprotection is effected by illumination of the substrate through, for example, a mask wherein the pattern has transparent regions with dimensions of, for example, less than 1 cm^2, 10^{-1} cm^2, 10^{-2} cm^2, 10^{-3} cm^2, 10^{-4} cm^2, 10^{-5} cm^2, 10^{-6} cm^2, 10^{-7} cm^2, 10^{-8} cm^2, or 10^{-10} cm^2. In a preferred embodiment, the regions are between about 10×10 μm and 500×500 μm. According to some embodiments, the masks are arranged to produce a checkerboard array of polymers, although any one of a variety of geometric configurations may be utilized.

1. Example

In one example of the invention, free amino groups were fluorescently labelled by treatment of the entire substrate surface with fluorescein isothiocyanate (FITC) after photodeprotection. Glass microscope slides were cleaned, aminated by treatment with 0.1% aminopropyltriethoxysilane in 95% ethanol, and incubated at 110° C. for 20 min. The aminated surface of the slide was then exposed to a 30 mM solution of the N-hydroxysuccinimide ester of NVOC-GABA (nitroveratryloxycarbonyl-τ-amino butyric acid) in DMF. The NVOC protecting group was photolytically removed by imaging the 365 nm output from a Hg arc lamp through a chrome on glass 100 μm checkerboard mask onto the substrate for 20 min at a power density of 12 mW/cm^2. The exposed surface was then treated with 1 mM FITC in DMF. The substrate surface was scanned in an epifluorescence microscope (Zeiss Axioskop 20) using 488 nm excitation from an argon ion laser (Spectra-Physics model 2025). The fluorescence emission above 520 nm was detected by a cooled photomultiplier (Hamamatsu 943-02) operated in a photon counting mode. Fluorescence intensity was translated into a color display with red in the highest intensity and black in the lowest intensity areas. The presence of a high-contrast fluorescent checkerboard pattern of 100×100 μm elements revealed that free amino groups were generated in specific regions by spatiallylocalized photodeprotection.

2. EXAMPLE

FIG. 2 is a flow chart illustrating another example of the invention. Carboxy-activated NVOC-leucine was allowed to react with an aminated substrate. The carboxy activated HOBT ester of leucine and other amino acids used in this synthesis was formed by mixing 0.25 mmol of the NVOC amino protected amino acid with 37 mg HOBT (1-hydroxybenzotriazole). 111 mg BOP (benzotriazolyl-n-oxy-tris (dimethylamino)-phosphoniumhexafluorophosphate) and 86 μl DIEA (diisopropylethylamine) in 2.5 ml DMF. The NVOC protecting group was removed by uniform illumination. Carboxy-activated NVOC-phenylalanine was coupled to the exposed amino groups for 2 hours at room temperature, and then washed with DMF and methylene chloride. Two unmasked cycles of photodeprotection and coupling with carboxy-activated NVOC-glycine were carried out. The surface was then illuminated through a chrome on glass 50 μl checkerboard pattern mask. Carboxy-activated Nα-tBOC-O-tButyl-L-tyrosine was then added. The entire surface was uniformly illuminated to photolyze the remaining NVOC groups. Finally, carboxy-activated NVOC-L-proline was added, the NVOC group was removed by illumination, and the t-BOC and t-butyl protecting groups were removed with TFA. After removal of the protecting groups, the surface consisted of a 50 μm checkerboard array of Tyr-Gly-Gly-Phe-Leu (YGGFL) (Seq. ID No:1) and Pro-Gly-Gly-Phe-Leu (PGGFL)(Seq. ID No:2).

B. Antibody Recognition

In one preferred embodiment the substrate is used to determine which of a plurality of amino acid sequences is recognized by an antibody of interest.

1. EXAMPLE

In one example, the array of pentapeptides in the example illustrated in FIG. 2 was probed with a mouse monoclonal antibody directed against β-endorphin. This antibody (called 3E7) is known to bind YGGFL and YGGFM (Seq. ID No:21) with nanomolar affinity and is discussed in Meo et al., *Proc. Natl. Acad. Sci. USA* (1983) 80:4084, which is incorporated by reference herein for all purposes. This antibody requires the amino terminal tyrosine for high affinity binding. The array of peptides formed as described in FIG. 2 was incubated with a 2 μg/ml mouse monoclonal antibody (3E7) known to recognize YGGFL. 3E7 does not bind PGGFL. A second incubation with fluoresceinated goat anti-mouse antibody labeled the regions that bound 3E7. The surface was scanned with an epi-fluorescence microscope. The results showed alternating bright and dark 50 μm squares indicating that YGGFL and PGGFL were synthesized in geometric array determined by the mask. A high contrast (>12:1 intensity ratio) fluorescence checkerboard image shows that (a) YGGFL and PGGFL were synthesized in alternate 50 μm squares, (b) YGGFL attached to the surface is accessible for binding to antibody 3E7. and (c) antibody 3E7 does not bind to PGGFL.

A three-dimensional representation of the fluorescence intensity data in a portion of the checkboard is shown in FIG. 3. This figure shows that the border between synthesis sites is sharp. The height of each spike in this display is linearly proportional to the integrated fluorescence intensity in a 2.5 μm pixel. The transition between PGGFL and YGGFL occurs within two spikes (5 μm). There is little variation in the fluorescence intensity of different YGGFL squares. The mean intensity of sixteen YGGFL synthesis sites was 2.03×10^5 counts and the standard deviation was 9.6×10^3 counts.

III. Synthesis

A. Reactor System

FIG. 4 schematically illustrates a device used to synthesize diverse polymer sequences on a substrate. The substrate, the area of synthesis, and the area for synthesis of each individual polymer could be of any size or shape. For example, squares, ellipsoids, rectangles, triangles, circles, or portions thereof, along with irregular geometric shapes may be utilized. Duplicate synthesis areas may also be applied to a single substrate for purposes of redundancy.

In one embodiment, the predefined regions on the substrate will have a surface area of between about 1 cm^2 and $10^{-10}cm^2$. In some embodiments the regions have areas of less than about 10^{-1} cm^2, 10^{-2} cm^2, $10^{-3}cm^2$, 10^{-4} cm^2, 10^{-5} cm^2, 10^{-6} cm^2, $10^{-7}cm^2$, 10^{-8} cm^2, 10^{-9} cm^2 or 10^{-10} cm^2. In a preferred embodiment, the regions are between about 10×10 μm.

in some embodiments a single substrate supports more than about 10 different monomer sequences and perferably more than about 100 different monomer sequences, although in some embodiments more than about 10^3, 10^4, 10^5, 10^6, 10^7, or 10^8 different sequences are provided on a substrate. Of course, within a region of the substrate in which a monomer sequence is synthesized, it is preferred that the monomer sequence be substantially pure. In some embodiments, regions of the substrate contain polymer sequences which are at least about 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98%, or 99% pure. The device includes an automated peptide synthesizer 401. The automated peptide synthesizer is a device which flows selected reagents through a flow cell 402 under the direction of a computer 404. In a preferred embodiment the synthesizer is an ABI Peptide Synthesizer, model no. 431A. The computer may be selected from a wide variety of computers or discrete logic including for, example, an IBM PC-AT or similar computer linked with appropriate internal control systems in the peptide synthesizer. The PC is provided with signals from the board computer indicative of, for example, the end of a coupling cycle.

Substrate 406 is mounted on the flow cell, forming a cavity between the substrate and the flow cell. Selected reagents flow through this cavity from the peptide synthesizer at selected times, forming an array of peptides on the face of the substrate in the cavity. Mounted above the substrate, and preferably in contact with the substrate is a mask 408. Mask 408 is transparent in selected regions to a selected wavelength of light and is opaque in other regions to the selected wavelength of light. The mask is illuminated with a light source 410 such as a UV light source. In one specific embodiment the light source 410 is a model no. 82420 made by Oriel. The mask is held and translated by an x-y-z translation stage 412 such as an x-y translation stage made by Newport Corp. The computer coordinates action of the peptide synthesizer, x-y translation stage, and light source. Of course, the invention may be used in some embodiments with translation of the substrate instead of the mask.

In operation, the substrate is mounted on the reactor cavity. The slide, with its surface protected by a suitable photo removable protective group, is exposed to light at selected locations by positioning the mask and illuminating the light source for a desired period of time (such as, for example, 1 sec to 60 min in the case of peptide synthesis). A selected peptide or other monomer/polymer is pumped through the reactor cavity by the peptide synthesizer for binding at the selected locations on the substrate. After a selected reaction time (such as about 1 sec to 300 min in the case of peptide reactions) of the monomer is washed from the system, the mask is appropriately repositioned or replaced, and the cycle is repeated. In most embodiments of the invention, reactions may be conducted at or near ambient temperature.

FIGS. 5*a* and 5*b* are flow charts of the software used in operation of the reactor system. At step 502 the peptide synthesis software is initialized. At step 504 the system calibrates positioners on the x-y translation stage and begins a main loop. At step 506 the system determines which, if any, of the function keys on the computer have been pressed. If F1 has been pressed, the system prompts the user for input of a desired synthesis process. If the user enters F2, the system allows a user to edit a file for a synthesis process at step 510. If the user enters F3 the system loads a process from a disk at step 512. If the user enters F4 the system saves an entered or edited process to disk at step 514. If the user selects F5 the current process is displayed at step 516 while selection of F6 starts the main portion of the program, i.e., the actual synthesis according to the selected process. If the user selects F7 the system displays the location of the synthesized peptides, while pressing F10 returns the user to the disk operating system.

FIG. 5*b* illustrates the synthesis step 518 in greater detail. The main loop of the program is started in which the system first moves the mask to a next position at step 526. During the main loop of the program, necessary chemicals flow through the reaction cell under the direction of the on-board computer in the peptide synthesizer. At step 528 the system then waits for an exposure command and, upon receipt of the exposure command exposes the substrate for a desired time at step 530. When an acknowledge of exposure complete is received at step 532 the system determines if the process is complete at step 534 and, if so, waits for additional keyboard input at step 536 and, thereafter, exits the perform synthesis process.

A computer program used for operation of the system described above is included as microfiche Appendix A (Copyright, 1990, Affymax Technologies N.V., all rights reserved). The program is written in Turbo C++ (Borland Int'l) and has been implemented in an IBM compatible system. The motor control software is adapted from software produced by Newport Corporation. It will be recognized that a large variety of programming languages could be utilized without departing from the scope of the invention herein. Certain calls are made to a graphics program in "Programmer Guide to PC and PS2 Video Systems" (Wilton, Microsoft Press, 1987), which is incorporated herein by reference for all purposes.

Alignment of the mask is achieved by one of two methods in preferred embodiments. In a first embodiment the system relies upon relative alignment of the various components, which is normally acceptable since x-y-z translation stages are capable of sufficient accuracy for the purposes herein. In alternative embodiments, alignment marks on the substrate are coupled to a CCD device for appropriate alignment.

According to some embodiments, pure reagents are not added at each step, or complete photolysis of the protective groups is not provided at each step. According to these embodiments, multiple products will be formed in each synthesis site. For example, if the monomers A and B are mixed during a synthesis step, A and B will bind to deprotected regions, roughly in proportion to their concentration in solution. Hence, a mixture of compounds will be formed in a synthesis region. A substrate formed with mixtures of compounds in various synthesis regions may be used to perform, for example, an initial screening of a large number of compounds, after which a smaller number of compounds in regions which exhibit high binding affinity are further screened. Similar results may be obtained by only partially photylizing a region, adding a first monomer, re-photylizing the same region, and exposing the region to a second monomer.

B. Binary Synthesis Strategy

In a light-directed chemical synthesis, the products formed depend on the pattern and order of masks, and on the order of reactants. To make a set of products there will in general be "n" possible masking schemes. In preferred embodiments of the invention herein a binary synthesis strategy-is utilized. The binary synthesis strategy is illustrated herein primarily with regard to a masking strategy, although it will be applicable to other polymer synthesis strategies such as the pin strategy, and the like.

In a binary synthesis strategy, the substrate is irradiated with a first mask, exposed to a first building block, irradiated with a second mask, exposed to a second building block, etc. Each combination of masked irradiation and exposure to a building block is referred to herein as a "cycle."

In a preferred binary masking scheme, the masks for each cycle allow irradiation of half of a region of interest on the substrate and protection of the remaining half of the region of interest. By "half" it is intended herein not to mean exactly one-half the region of interest, but instead a large fraction of the region of interest such as from about 30 to 70 percent of the region of interest. It will be understood that the entire masking scheme need not take a binary form; instead non-binary cycles may be introduced as desired between binary cycles.

In preferred embodiments of the binary masking scheme, a given cycle illuminates only about half of the region which was illuminated in a previous cycle, while protecting the remaining half of the illuminated portion from the previous cycle. Conversely, in such preferred embodiments, a given cycle illuminates half of the region which was protected in the previous cycle and protects half the region which was protected in a previous cycle.

The synthesis strategy is most readily illustrated and handled in matrix notation. At each synthesis site, the determination of whether to add a given monomer is a binary process. Therefore, each product element P_j is given by the dot product of two vectors, a chemical reactant vector, e.g., C=[A,B,C,D], and a binary vector σ_j. Inspection of the products in the example below for a four-step synthesis, shows that in one four-step synthesis σ_1=[1,0,1,0], σ_2=[1,0,0,1], σ_3=[0.1.1.0], and σ_4=[0.1,0.1], where a 1 indicates illumination and a 0 indicates protection. Therefore, it becomes possible to build a "switch matrix" S from the column vectors σ_j (j=1,k where k is the number of products).

$$
\begin{array}{cccc}
\sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \\
1 & 1 & 0 & 0 \\
S=0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1
\end{array}
$$

The outcome P of a synthesis is simply P=CS, the product of the chemical reactant matrix and the switch matrix.

The switch matrix for an n-cycle synthesis yielding k products has n rows and k columns. An important attribute of S is that each row specifies a mask. A two-dimensional mask m_j for the jth chemical step of a synthesis is obtained directly from the jth row of S by placing the elements s_{j1}, . . . s_{jk} into, for example, a square format. The particular arrangement below provides a square format, although linear or other arrangements may be utilized.

$$S=\begin{matrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{matrix} \quad m_j=\begin{matrix} s_{j1} & s_{j2} \\ s_{j3} & s_{j4} \end{matrix}$$

Of course, compounds formed in a light-activated synthesis can be positioned in any defined geometric array. A square or rectangular matrix is convenient but not required. The rows of the switch matrix may be transformed into any convenient array as long as equivalent transformations are used for each row.

For example, the masks in the four-step synthesis below are then denoted by:

$$m_1=\begin{matrix} 1 & 1 \\ 0 & 0 \end{matrix}\; m_2=\begin{matrix} 0 & 0 \\ 1 & 1 \end{matrix}\; m_3=\begin{matrix} 1 & 0 \\ 1 & 0 \end{matrix}\; m_4=\begin{matrix} 0 & 1 \\ 0 & 1 \end{matrix}$$

where 1 denotes illumination (activation) and 0 denotes no illumination.

The matrix representation is used to generate a desired set of products and product maps in preferred embodiments. Each compound is defined by the product of the chemical vector and a particular switch vector. Therefore, for each synthesis address, one simply saves the switch vector, assembles all of them into a switch matrix, and extracts each of the rows to form the masks.

In some cases, particular product distributions or a maximal number of products are desired. For example, for C=[A,B,C,D], any switch vector (σ_j) consists of four bits. Sixteen four-bit vectors exist. Hence, a maximum of 16 different products can be made by sequential addition of the reagents [A,B,C,D]. These 16 column vectors can be assembled in 16! different ways to form a switch matrix. The order of the column vectors defines the masking patterns, and, therefore, the spatial ordering of products but not their makeup. One ordering of these columns gives the following switch matrix (in which "null" (θ) additions are included in brackets for the sake of completeness, although such null additions are elsewhere ignored herein):

$$S=\begin{array}{lccccccccccccccll}
\sigma 1 & & & & & & & & & & & & & & & \sigma_{16} & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A \\
[0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1] & \phi \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & B \\
[0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1] & \phi \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & C \\
[0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1] & \phi \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & D \\
[0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1] & \phi
\end{array}$$

The columns of S according to this aspect of the invention are the binary representations of the numbers 15 to 0. The sixteen products of this binary synthesis are ABCD, ABC, ABD, AB, ACD, AC, AD, A, BCD, BC, BD, B, CD, C, D, and θ (null). Also note that each of the switch vectors from the four-step synthesis masks above (and hence the synthesis products) are present in the four bit binary switch matrix. (See columns 6, 7, 10, and 11)

This synthesis procedure provides an easy way for mapping the completed products. The products in the various locations on the substrate are simply defined by the columns of the switch matrix (the first column indicating, for example, that the product ABCD will be present in the upper left-hand location of the substrate). Furthermore, if only selected desired products are to be made, the mask sequence can be derived by extracting the columns with the desired sequences. For example, to form the product set ABCD, ABD, ACD, AD, BCD, BD, CD, and D, the masks are formed by use of a switch matrix with only the 1st, 3rd, 5th, 7th, 9th, 11th, 13th, and 15th columns arranged into the switch matrix:

$$S=\begin{matrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{matrix}$$

To form all of the polymers of length 4, the reactant matrix [ABCDABCDABCDABCD]is used. The switch matrix will be formed from a matrix of the binary numbers from 0 to 2^{16} arranged in columns. The columns having four monomers are than selected and arranged into a switch matrix. Therefore, it is seen that the binary switch matrix in general will provide a representation of all the products which can be made from an n-step synthesis, from which the desired products are then extracted.

The rows of the binary switch matrix will, in preferred embodiments, have the property that each masking step illuminates half of the synthesis area. Each masking step also factors the preceding masking step; that is, half of the region that was illuminated in the preceding step is again illuminated, whereas the other half is not. Half of the region that was unilluminated in the preceding step is also illuminated, whereas the other half is not. Thus, masking is recursive. The masks are constructed, as described previously, by extracting the elements of each row and placing them in a square array. For example, the four masks in S for a four-step synthesis are:

$$m_1=\begin{matrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{matrix}\; m_2=\begin{matrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{matrix}$$

$$m_3=\begin{matrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{matrix}\; m_4=\begin{matrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{matrix}$$

The recursive factoring of masks allows the products of a light-directed synthesis to be represented by a polynomial. (Some light activated syntheses can only be denoted by irreducible, i.e., prime polynomials.) For example, the polynomial corresponding to the top synthesis of FIG. 9a (discussed below) is

$$P=(A+B)(C+D)$$

A reaction polynomial may be expanded as though it were an algebraic expression, provided that the order of joining of reactants X_1 and X_2 is preserved ($X_1X_2 \neq X_2X_1$), i.e., the products are not commutative. The product then is AC+AD+BC+BD. The polynomial explicitly specifies the reactants and implicitly specifies the mask for each step. Each pair of parentheses demarcates a round of synthesis. The chemical reactants of a round (e.g., A and B) react at nonoverlapping sites and hence cannot combine with one other. The synthesis area is divided equally amongst the elements of a round (e.g., A is directed to one-half of the area and B to the other half). Hence, the masks for a round (e.g., the masks m_A and mB) are orthogonal and form an orthonormal set. The polynomial notation also signifies that each element in a round is to be joined to each element of the next round (e.g., A with C, A with D, B with C, and B with D). This is accomplished by having m_C overlap m_A an m_B equally, and likewise for m_D. Because C and D are elements of a round, m_C and m_D are orthogonal to each other and form an orthonormal set.

The polynomial representation of the binary synthesis described above, in which 16 products are made from 4 reactants, is

$$P=(A+\theta)(B+\theta)\ (C+\theta)\ (D+\theta)$$

which gives ABCD, ABC, ABD, AB, ACD, AC, AD, A, BCD, BC, BD, B, CD, C, D, and θ when expanded (with the rule that θX=X and Xθ=X, and remembering that joining is ordered). In a binary synthesis, each round contains one reactant and one null (denoted by θ). Half of the synthesis area receives the reactant and the other half receives nothing. Each mask overlaps every other mask equally.

Binary rounds and non-binary rounds can be interspersed as desired, as in

$$P=(A+\theta)(B)(C+D+\theta)(E+F+G)$$

The 18 compounds formed are ABCE, ABCF, ABCG, ABDE, ABDF, ABDG, ABE, ABF, ABG, BCE, BCF, BCG, BDE, BDF, BDG, BE, BF, and BG. The switch matrix S for this 7-step synthesis is

$$S=\begin{array}{cccccccccccccccccc}
1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0\\
0&0&0&1&1&1&0&0&0&0&0&0&1&1&1&0&0&0\\
1&0&0&1&0&0&1&0&0&1&0&0&1&0&0&1&0&0\\
0&1&0&0&1&0&0&1&0&0&1&0&0&1&0&0&1&0\\
0&0&1&0&0&1&0&0&1&0&0&1&0&0&1&0&0&1
\end{array}$$

The round denoted by (B) places B in all products because the reaction area was uniformly activated (the mask for B consisted entirely of 1's).

The number of compounds k formed in a synthesis consisting of r rounds, in which the ith round has b_i chemical reactants and z_i nulls, is

$$k=\Sigma(b_i+z_i)$$

and the number of chemical steps n is

$$n=\Sigma b_i$$

The number of compounds synthesized when b=a and z=0 in all rounds is $a^{n/a}$, compared with 2^n for a binary synthesis. For n=20 and a=5, 625 compounds (all tetrameros) would be formed, compared with 1.049×10^6 compounds in a binary synthesis with the same number of chemical steps.

It should also be noted that rounds in a polynomial can be nested, as in

$$(A+(B+\theta)(C+\theta)(D+\theta)$$

The products are AD, BCD, BD, CD, D, A, BC, B, C, and θ.

Binary syntheses are attractive for two reasons. First, they generate the maximal number of products (2^n) for a given number of chemical steps (n). For four reactants, 16 compounds are formed in the binary synthesis, whereas only 4 are made when each round has two reactants. A 10-step binary synthesis yields 1,024 compounds, and a 20-step synthesis yields 1,048,576. Second, products formed in a binary synthesis are a complete nested set with lengths ranging from 0 to n. All compounds that can be formed by deleting one or more units from the longest product (the n-mer) are present. Contained within the binary set are the smaller sets that would be formed from the same reactants using any other set of masks (e.g., AC, AD, BC, and BD formed in the synthesis shown in FIG. 6 are present in the set of 16 formed by the binary synthesis). In some cases, however, the experimentally achievable spatial resolution may not suffice to accommodate all the compounds formed. Therefore, practical limitations may require one to select a particular subset of the possible switch vectors for a given synthesis.

1. EXAMPLE

FIG. 6 illustrates a synthesis with binary masking scheme. The binary masking scheme provides the greatest number of sequences for a given number of cycles. According to this embodiment, a mask m1 allows illumination of half of the substrate. The substrate is then exposed to the building block A, which binds at the illuminated regions.

Thereafter, the mask m2 allows illumination of half of the previously illuminated region, while protecting half of the previously illuminated region. The building block B is then added, which binds at the illuminated regions from m2.

The process continues with masks m3, m4, and m5, resulting in the product array shown in the bottom portion of the figure. The process generates 32 (2 raised to the power of the number of monomers) sequences with 5 (the number of monomers) cycles.

2. EXAMPLE

FIG. 7 illustrates another preferred binary masking scheme which is referred to herein as the gray code masking scheme. According to this embodiment, the masks m1 to m5 are selected such that a side of any given synthesis region is defined by the edge of only one mask. The site at which the sequence BCDE is formed, for example, has its right edge defined by m5 and its left side formed by mask m4 (and no other mask is aligned on the sides of this site). Accordingly, problems created by misalignment, diffusion of light under the mask and the like will be minimized.

3. EXAMPLE

FIG. 8 illustrates another binary masking scheme. According to this scheme, referred to herein as a modified gray code masking scheme, the number of masks needed is minimized. For example, the mask m2 could be the same mask as m1 and simply translated laterally. Similarly, the mask m4 could be the same as mask m3 and simply translated laterally.

4. EXAMPLE

A four-step synthesis is shown in FIG. 9a. The reactants are the ordered set {A,B,C,D}. In the first cycle, illumination through m_1 activates the upper half of the synthesis area. Building block A is then added to give the distribution 602. Illumination through mask m_2 (which activates the lower half), followed by addition of B yields the next intermediate distribution 604. C is added after illumination through m_3 (which activates the left half) giving the distribution 604, and D after illumination through m_4 (which activates the right half), to yield the final product pattern 608 (AC,AD, BC,BD).

5. EXAMPLE

The above masking strategy for the synthesis may be extended for all 400 dipeptides from the 20 naturally occurring amino acids as shown in FIG. 9*b*. The synthesis consists of two rounds. with 20 photolysis and chemical coupling cycles per round. In the first cycle of round 1. mask 1 activates 1/20th of the substrate for coupling with the first of 20 amino acids. Nineteen subsequent illumination/coupling cycles in round 1 yield a substrate consisting of 20 rectangular stripes each bearing a distinct member of the 20 amino acids. The masks of round 2 are perpendicular to round 1 masks and therefore a single illumination/coupling cycle in round 2 yields 20 dipeptides. The 20 illumination/coupling cycles of round 2 complete the synthesis of the 400 dipeptides.

6. EXAMPLE

The power of the binary masking strategy can be appreciated by the outcome of a 10-step synthesis that produced 1.024 peptides. The polynomial expression for this 10-step binary synthesis was:

$$(f+\theta)(Y+\theta)(G+\theta)(A+\theta)(G+\theta)(T+\theta)\ (F+\theta)(L+\theta)(S+\theta)(F+\theta)$$

Each peptide occupied a 400×400 µm square. A 32x32 peptide array (1.024 peptides, including the null peptide and 10 peptides of l=1. and a limited number of duplicates) was clearly evident in a fluorescence scan following side group deprotection and treatment with the antibody 3E7 and fluorescinated antibody. Each synthesis site was a 400×400 µm square.

The scan showed a range of fluorescence intensities. from a background value of 3.300 counts to 22.400 counts in the brightest square (x=20. y=9). Only 15 compounds exhibited an intensity greater than 12.300 counts. The median value of the array was 4.800 counts.

The identity of each peptide in the array could be determined from its x and y coordinates (each range from 0 to 31) and the map of FIG. 10. The chemical units at positions 2. 5. 6. 9. and 10 are specified by the y coordinate and those at positions 1. 3. 4. 7. 8 by the x coordinate. All but one of the peptides was shorter than 10 residues. For example. the peptide at x=12 and y=3 is YGAGF (SEQ. ID No:3) (positions 1. 6. 8. 9. and 10 are nulls). YGAFLS (SEQ. ID No:4). the brightest element of the array. is at x=20 and y=9.

It is often desirable to deduce a binding affinity of a given peptide from the measured fluorescence intensity. Conceptually. the simplest case is one in which a single peptide binds to a univalent antibody molecule. The fluorescence scan is carried out after the slide is washed with buffer for a defined time. The order of fluorescence intensities is then a measure primarily of the relative dissociation rates of the antibody-peptide complexes. If the on-rate constants are the same (e.g., if they are diffusion-controlled), the order of fluorescence intensities will correspond to the order of binding affinities. However. the situation is sometimes more complex because a bivalent primary antibody and a bivalent secondary antibody are used. The density of peptides in a synthesis area corresponded to a mean separation of ~7 nm. which would allow multivalent antibody-peptide interactions. Hence. fluorescence intensities obtained according to the method herein will often be a qualitative indicator of binding affinity.

Another important consideration is the fidelity of synthesis. Deletions are produced by incomplete photodeprotection or incomplete coupling. The coupling yield per cycle in these experiments is typically between 85% and 95%. Implementing the switch matrix by masking is imperfect because of light diffraction. internal reflection, and scattering. Consequently, stowaways (chemical units that should not be on board) arise by unintended illumination of regions that should be dark. A binary synthesis array contains many of the controls needed to assess the fidelity of a synthesis. For example. the fluorescence signal from a synthesis area nominally containing a tetrapeptide ABCD could come from a tripeptide deletion impurity such as ACD. Such an artifact would be ruled out by the finding that the fluorescence intensity of the ACD-site is less than that of the ABCD site.

The fifteen most highly labelled peptides in the array obtained with the synthesis of 1.024 peptides described above. were YGAFLS (SEQ. ID No:5). YGAFS (SEQ. ID No:6). YGAFL (SEQ. ID No:7). YGGFLS (SEQ. ID No:8). YGAF (SEQ. ID No:8). YGALS (SEQ. ID No:9). YGGFS (SEQ. ID No:10). YGAL (SEQ. ID No:11). YGAFLF (SEQ. ID No:12), YGAF (SEQ. ID No:13). YGAFF (SEQ. ID No:14), YGGLS (SEQ. ID No:15). YGGFL (SEQ. ID No:16), SEQ. ID No:17), and YGAFLSF (SEQ. I fifteen begin with YG. which agrees with previous work showing that an amino-terminal tyrosine is a key determinant of binding. Residue 3 of this set is either A or G. and residue 4 is either F or L. The exclusion of S and T from these positions is clear cut. The finding that the preferred sequence is YG (A/G) (F/L) fits nicely with the outcome of a study in which a very large library of peptides on phage generated by recombinant DNA methods was screened for binding to antibody 3E7 (see Cwirla et al., *Proc. Natl. Acad. Sci.* USA. (1990) 87:6378. incorporated herein by reference). Additional binary syntheses based on leads from peptides on phage experiments show that YGAFMQ (SEQ. ID No:18), YGAFM (SEQ. ID No:19), and YGAFQ (SEQ. ID No:20) give stronger fluorescence signals than does YGGFM. the immunogen used to obtain antibody 3E7.

Variations on the above masking strategy will be valuable in certain circumstances. For example. if a "kernel" sequence of interest consists of PQR separated from XYZ and that the aim is to synthesize peptides in which these units are separated by a variable number of different residues. then the kernel can be placed in each peptide by using a mask that has 1's everywhere. The polynomial representation of a suitable synthesis is:

$$(P)(Q)(R)(A+\theta)(B+\theta)(C+\theta)(D+\theta)(X)(Y)(Z)$$

Sixteen peptides will be formed, ranging in length from the 6-mer PQRXYZ to the 10-mer PQRABCDXYZ.

Several other masking strategies will also find value in selected circumstances. By using a particular mask more than once. two or more reactants will appear in the same set of products. For example, suppose that the mask for an 8-step synthesis is

A	11110000
B	00001111
C	11001100
D	00110011
E	10101010
F	01010101
G	11110000
H	00001111

The products are ACEG, ACFG, ADEG. ADFG, BCEH. BCFH, BDEH. and BDFH. A and G always appear together because their additions were directed by the same mask, and likewise for B and H.

C. Linker Selection

According to preferred embodiments the linker molecules used as an intermediary between the synthesized polymers and the substrate are selected for optimum length and/or type for improved binding interaction with a receptor. According to this aspect of the invention diverse linkers of varying length and/or type are synthesized for subsequent attachment of a ligand. Through variations in the length and type of linker, it becomes possible to optimize the binding interaction between an immobilized ligand and its receptor.

The degree of binding between a ligand (peptide, inhibitor, hapten, drug, etc.) and its receptor (enzyme, antibody, etc.) when one of the partners is immobilized on to a substrate will in some embodiments depend on the accessibility of the receptor in solution to the immobilized ligand. The accessibility in turn will depend on the length and/or type of linker molecule employed to immobilize one of the partners. Preferred embodiments of the invention therefore employ the ULSIPS™ technology described herein to generate an array of, preferably, inactive or inert linkers of varying length and/or type, using photochemical protecting groups to selectively expose different regions of the substrate and to build upon chemically-active groups.

In the simplest embodiment of this concept, the same unit is attached to the substrate in varying multiples or lengths in known locations on the substrate via VLSIPS™ techniques to generate an array of polymers of varying length. A single ligand (peptide, drug, hapten, etc.) is attached to each of them, and an assay is performed with the binding site to evaluate the degree of binding with a receptor that is known to bind to the ligand. In cases where the linker length impacts the ability of the receptor to bind to the ligand, varying levels of binding will be observed. In general, the linker which provides the highest binding will then be used to assay other ligands synthesized in accordance with the techniques herein.

According to other embodiments the binding between a single ligand/receptor pair is evaluated for linkers of diverse monomer sequence. According to these embodiments, the linkers are synthesized in an array in accordance with the techniques herein and have different monomer sequence (and, optionally, different lengths). Thereafter, all of the linker molecules are provided with a ligand known to have at least some binding affinity for a given receptor. The given receptor is then exposed to the ligand and binding affinity is deduced. Linker molecules which provide adequate binding between the ligand and receptor are then utilized in screening studies.

D. Protecting Groups

As discussed above, selectively removable protecting groups allow creation of well defined areas of substrate surface having differing reactivities. Preferably, the protecting groups are selectively removed from the surface by applying a specific activator, such as electromagnetic radiation of a specific wavelength and intensity. More preferably, the specific activator exposes selected areas of surface to remove the protecting groups in the exposed areas.

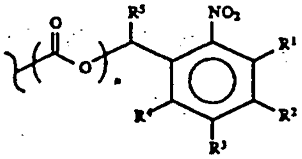
Protecting groups of the present invention are used in conjunction with solid phase oligomer syntheses, such as peptide syntheses using natural or unnatural amino acids, nucleotide syntheses using deoxyribonucleic and ribonucleic acids, oligosaccharide syntheses, and the like. In addition to protecting the substrate surface from unwanted reaction, the protecting groups block a reactive end of the monomer to prevent self-polymerization. For instance, attachment of a protecting group to the amino terminus of an activated amino acid, such as an N-hydroxysuccinimide-activated ester of the amino acid, prevents the amino terminus of one monomer from reacting with the activated ester portion of another during peptide synthesis. Alternatively, the protecting group may be attached to the carboxyl group of an amino acid to prevent reaction at this site. Most protecting groups can be attached to either the amino or the carboxyl group of an amino acid, and the nature of the chemical synthesis will dictate which reactive group will require a protecting group. Analogously, attachment of a protecting group to the 5'-hydroxyl group of a nucleoside during synthesis using for example, phosphate-triester coupling chemistry, prevents the 5'-hydroxyl of one nucleoside from reacting with the 3'-activated phosphate-triester of another.

Regardless of the specific use, protecting groups are employed to protect a moiety on a molecule from reacting with another reagent. Protecting groups of the present invention have the following characteristics: they prevent selected reagents from modifying the group to which they are attached; they are stable (that is, they remain attached to the molecule) to the synthesis reaction conditions; they are removable under conditions that do not adversely affect the remaining structure; and once removed, do not react appreciably with the surface or surface-bound oligomer. The selection of a suitable protecting group will depend, of course, on the chemical nature of the monomer unit and oligomer, as well as the specific reagents they are to protect against.

In a preferred embodiment, the protecting groups are photoactivatable. The properties and uses of photoreactive protecting compounds have been reviewed. See, McCray et al., *Ann. Rev. of Biophys. and Biophys. Chem.* (1989) 18:239–270, which is incorporated herein by reference. Preferably, the photosensitive protecting groups will be removable by radiation in the ultraviolet (UV) or visible portion of the electromagnetic spectrum. More preferably, the protecting groups will be removable by radiation in the near UV or visible portion of the spectrum. In some embodiments, however, activation may be performed by other methods such as localized heating, electron beam lithography, laser pumping, oxidation or reduction with microelectrodes, and the like. Sulfonyl compounds are suitable reactive groups for electron beam lithography. Oxidative or reductive removal is accomplished by exposure of the protecting group to an electric current source, preferably using microelectrodes directed to the predefined regions of the surface which are desired for activation. Other methods may be used in light of this disclosure.

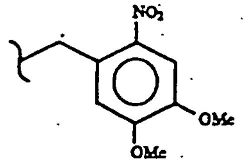
Many, although not all, of the photoremovable protecting groups will be aromatic compounds that absorb near-UV and visible radiation. Suitable photoremovable protecting groups are described in, for example, McCray et al., Patchornik, *J. Amer. Chem. Soc.* (1970) 92:6333, and Amit et al., *J. Org. Chem.* (1974) 39:192, which are incorporated herein by reference.

A preferred class of photoremovable protecting groups has the general formula:

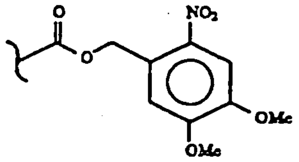


where R^1, R^2, R^3, and R^4 independently are a hydrogen atom, a lower alkyl, aryl, benzyl, halogen, hydroxyl, alkoxyl, thiol, thioether, amino, nitro, carboxyl, formate, formamido or phosphido group, or adjacent substituents (i.e., R^1-R^2, R^2-R^3, R^3-R^4) are substituted oxygen groups that together form an cyclic acetal or ketal; R^5 is a hydrogen atom, a alkoxyl, alkyl, hydrogen, halo, aryl, or alkenyl group, and n=0 or 1.

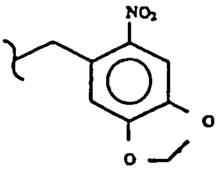
A preferred protecting group, 6-nitroveratryl (NV), which is used for protecting the carboxyl terminus of an amino acid or the hydroxyl group of a nucleotide, for example, is formed when R^2 and R^3 are each a methoxygroup, R^1, R^4 and R^5 are each a hydrogen atom, and n=0:



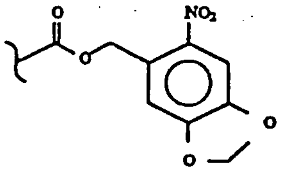
A preferred protecting group, 6-nitroveratryloxycarbonyl (NVOC), which is used to protect the amino terminus of an amino acid, for example, is formed when R^2 and R^3 are each a methoxy group, R^1, R^4 and R^5 are each a hydrogen atom, and n=1:



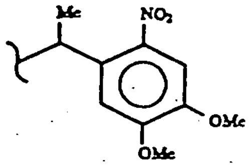
Another preferred protecting group, 6-nitropiperonyl (NP), which is used for protecting the carboxyl terminus of an amino acid or the hydroxyl group of a nucleotide, for example, is formed when R^2 and R^3 together form a methylene acetal, R^1, R^4 and R^5 are each a hydrogen atom, and n=0:



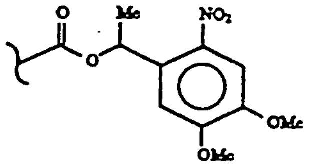
Another preferred protecting group, 6-nitropiperonyloxycarbonyl (NPOC), which is used to protect the amino terminus of an amino acid, for example, is formed when R^2 and R^3 together form a methylene acetal, R^1, R^4 and R^5 are each a hydrogen atom, and n=1:



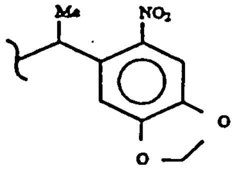
A most preferred protecting group, methyl-6-nitroveratryl (MeNV), which is used for protecting the carboxyl terminus of an amino acid or the hydroxyl group of a nucleotide, for example, is formed when R^2 and R^3 are each a methoxy group, R^1 and R^4 are each a hydrogen atom, R^5 is a methyl group, and n=0:



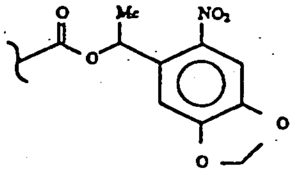
Another most preferred protecting group, methyl-6-nitroveratryloxycarbonyl (MeNVOC), which is used to protect the amino terminus of an amino acid, for example, is formed when R^2 and R^3 are each a methoxy group, R^1 and R^4 are each a hydrogen atom, R^5 is a methyl group, and n=1:



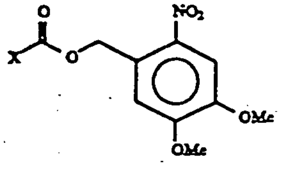
Another most preferred protecting group, methyl-6-nitropiperonyl (MeNP), which is used for protecting the carboxyl terminus of an amino acid or the hydroxyl group of a nucleotide, for example, is formed when R^2 and R^3 together form a methylene acetal, R^1 and R^4 are each a hydrogen atom, R^5 is a methyl group, and n=0:



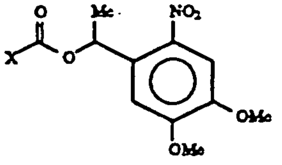
Another most preferred protecting group, methyl-6-nitropiperonyloxycarbonyl (MeNPOC), which is used to protect the amino terminus of an amino acid, for example, is formed when R^2 and R^3 together form a methylene acetal, R^1 and R^4 are each a hydrogen atom, R^5 is a methyl group, and n=1:



A protected amino acid having a photoactivatable oxycarbonyl protecting group, such NVOC or NPOC or their corresponding methyl derivatives, MeNVOC or MeNPOC, respectively, on the amino terminus is formed by acylating the amine of the amino acid with an activated oxycarbonyl ester of the protecting group. Examples of activated oxycarbonyl esters of NVOC and MeNVOC have the general formula:



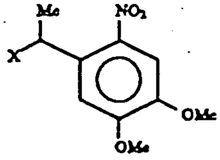
NVOC-X



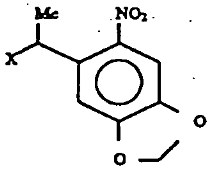
MeNVOC-X

where X is halogen, mixed anhydride, phenoxy, p-nitrophenoxy, N-hydroxysuccinimide, and the like.

A protected amino acid or nucleotide having a photoactivatable protecting group, such as NV or NP or their corresponding methyl derivatives, MeNV or MeNP, respectively, on the carboxy terminus of the amino acid or 5'-hydroxy terminus of the nucleotide, is formed by acylating the carboxy terminus or 5'-OH with an activated benzyl derivative of the protecting group. Examples of activated benzyl derivatives of MeNV and MeNP have the general formula:



MeNV-X

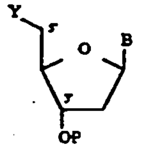


MeNP-X

where X is halogen, hydroxyl, tosyl, mesyl, trifluoromethyl, diazo, azido, and the like.

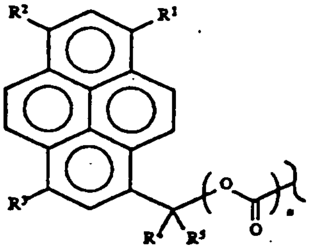
Another method for generating protected monomers is to react the benzylic alcohol derivative of the protecting group with an activated ester of the monomer. For example, to protect the carboxyl terminus of an amino acid, an activated ester of the amino acid is reacted with the alcohol derivative of the protecting group, such as 6-nitroveratrol (NVOH). Examples of activated esters suitable for such uses include halo-formate, mixed anhydride, imidazoyl formater acyl halide, and also includes formation of the activated ester in situ the use of common reagents such as DCC and the like. See Atherton et al. for other examples of activated esters.

A further method for generating protected monomers is to react the benzylic alcohol derivative of the protecting group with an activated carbon of the monomer. For example, to protect the 5'-hydroxyl group of a nucleic acid, a derivative having a 5'-activated carbon is reacted with the alcohol derivative of the protecting group, such as methyl-6-nitropiperonol (MePyROH). Examples of nucleotides having activating groups attached to the 5'-hydroxyl group have the general formula:



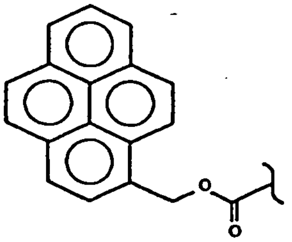
where Y is a halogen atom, a tosyl, mesyl, trifluoromethyl, azido, or diazo group, and the like.

Another class of preferred photochemical protecting groups has the formula:

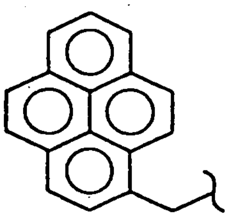


where R^1, R^2, and R^3 independently are a hydrogen atom, a lower alkyl, aryl, benzyl, halogen, hydroxyl, alkoxyl, thiol, thioether, amino, nitro, carboxyl, formate, formamido, sulfanates, sulfido or phosphido group, R^4 and R^5 independently are a hydrogen atom, an alkoxy, alkyl, halo, aryl, hydrogen, or alkenyl group, and n=0 or 1.

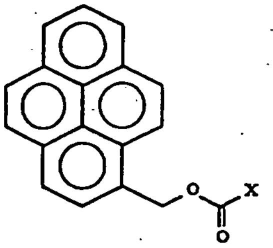
A preferred protecting group, 1-pyrenylmethyloxycarbonyl (PyROC), which is used to protect the amino terminus of an amino acid, for example, is formed when R^1 through R^5 are each a hydrogen atom and n=1:



Another preferred protecting group, 1-pyrenylmethyl (PyR), which is used for protecting the carboxy terminus of an amino acid or the hydroxyl group of a nucleotide, for example, is formed when R^1 through R^5 are each a hydrogen atom and n=0:

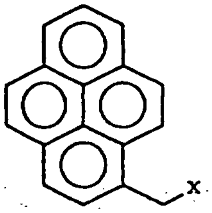


An amino acid having a pyrenylmethyloxycarbonyl protecting group on its amino terminus is formed by acylation of the free amine of amino acid with an activated oxycarbonyl ester of the pyrenyl protecting group. Examples of activated oxycarbonyl esters of PyROC have the general formula:



where X is halogen, or mixed anhydride, p-nitrophenoxy, or N-hydroxysuccinimide group, and the like.

A protected amino acid or nucleotide having a photoactivatable protecting group, such as PyR, on the carboxy terminus of the amino acid or 5'-hydroxy terminus of the nucleic acid, respectively, is formed by acylating the carboxy terminus or 5'-OH with an activated pyrenylmethyl derivative of the protecting group. Examples of activated pyrenylmethyl derivatives of PyR have the general formula:



where X is a halogen atom, a hydroxyl, diazo, or azido group, and the like.

Another method of generating protected monomers is to react the pyrenylmethyl alcohol moiety of the protecting group with an activated ester of the monomer. For example, an activated ester of an amino acid can be reacted with the alcohol derivative of the protecting group, such as pyrenylmethyl alcohol (PyROH), to form the protected derivative of the carboxy terminus of the amino acid. Examples of activated esters include halo-formate, mixed anhydride, imidazoyl formate, acyl halide, and also includes formation of the activated ester in situ and the use of common reagents such as DCC and the like.

Clearly, many photosensitive protecting groups are suitable for use in the present invention.

In preferred embodiments, the substrate is irradiated to remove the photoremovable protecting groups and create regions having free reactive moieties and side products resulting from the protecting group. The removal rate of the protecting groups depends on the wavelength and intensity of the incident radiation, as well as the physical and chemical properties of the protecting group itself. Preferred protecting groups are removed at a faster rate and with a lower intensity of radiation. For example, at a given set of conditions, MeNVOC and MeNPOC are photolytically removed from the N-terminus of a peptide chain faster than their unsubstituted parent compounds, NVOC and NPOC, respectively.

Removal of the protecting group is accomplished by irradiation to liberate the reactive group and degradation products derived from the protecting group. Not wishing to be bound by theory, it is believed that irradiation of an NVOC- and MeNVOC-protected oligomers occurs by the following reaction schemes:

NVOC-AA→3,4-dimethoxy-6-nitrosobenzaldehyde+CO_2+AA

MeNVOC-AA→3,4-dimethoxy-6-nitrosoacetophenone+CO_2+AA

where AA represents the N-terminus of the amino acid oligomer.

Along with the unprotected amino acid, other products are liberated into solution: carbon dioxide and a 2,3-dimethoxy-6-nitrosophenylcarbonyl compound, which can react with nucleophilic portions of the oligomer to form unwanted secondary reactions. In the case of an NVOC-protected amino acid, the degradation product is a nitrosobenzaldehyde, while the degradation product for the other is a nitrosophenyl ketone. For instance, it is believed that the product aldehyde from NVOC degradation reacts with free amines to form a Schiff base (imine) that affects the remaining polymer synthesis. Preferred photoremovable protecting groups react slowly or reversibly with the oligomer on the support.

Again not wishing to be bound by theory, it is believed that the product ketone from irradiation of a MeNVOC-protected oligomer reacts at a slower rate with nucleophiles on the oligomer than the product aldehyde from irradiation of the same NVOC-protected oligomer. Although not unambiguously determined, it is believed that this difference in reaction rate is due to the difference in general reactivity between aldehyde and ketones towards nucleophiles due to steric and electronic effects.

The photoremovable protecting groups of the present invention are readily removed. For example, the photolysis of N-protected L-phenylalanine in solution and having different photoremovable protecting groups was analyzed, and the results are presented in the following table:

TABLE

Photolysis of Protected L-Phe—OH

	$t_{1/2}$ in seconds			
Solvent	NBOC	NVOC	MeNVOC	MeNPOC
Dioxane	1288	110	24	19
5 mM H_2SO_4/Dioxane	1575	98	33	22

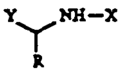
The half life, t1/2, is the time in seconds required to remove 50% of the starting amount of protecting group. NBOC is the 6-nitrobenzyloxycarbonyl group, NVOC is the 6-nitroveratryloxycarbonyl group, MeNVOC is the methyl-6-nitroveratryloxycarbonyl group, and MeNPOC is the methyl-6-nitropiperonyloxycarbonyl group. The photolysis was carried out in the indicated solvent with 362/364 nm-wavelength irradiation having an intensity of 10 mW/cm^2, and the concentration of each protected phenylalanine was 0.10 mM.

The table shows that deprotection of NVOC-, MeNVOC-, and MeNPOC-protected phenylalanine proceeded faster than the deprotection of NBOC. Furthermore, it shows that the deprotection of the two derivatives that are substituted on the benzylic carbon, MeNVOC and MeNPOC, were photolyzed at the highest rates in both dioxane and acidified dioxane.

1. Use of Photoremovable Groups During Solid-Phase Synthesis of Peptides

The formation of peptides on a solid-phase support requires the stepwise attachment of an amino acid to a substrate-bound growing chain. In order to prevent unwanted polymerization of the monomeric amino acid under the reaction conditions, protection of the amino terminus of the amino acid is required. After the monomer is coupled to the end of the peptide, the N-terminal protecting group is removed, and another amino acid is coupled to the chain. This cycle of coupling and deprotecting is continued for each amino acid in the peptide sequence. See Merrifield, *J. Am. Chem. Soc.* (1963) 85:2149, and Atherton et al., "Solid Phase Peptide Synthesis" 1989, IRL Press, London, both incorporated herein by reference for all purposes. As described above, the use of a photoremovable protecting group allows removal of selected portions of the substrate surface, via patterned irradiation, during the deprotection cycle of the solid phase synthesis. This selectively allows spatial control of the synthesis—the next amino acid is coupled only to the irradiated areas.

In one embodiment, the photoremovable protecting groups of the present invention are attached to an activated ester of an amino acid at the amino terminus:



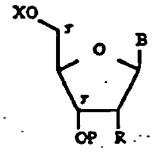
where R is the side chain of a natural or unnatural amino acid, X is a photoremovable protecting group, and Y is an activated carboxylic acid derivative. The photoremovable protecting group, X, is preferably NVOC, NPOC, PyROC, MeNVOC, MeNPOC, and the like as discussed above. The activated ester, Y, is preferably a reactive derivative having a high coupling efficiency, such as an acyl halide, mixed anhydride, N-hydroxysuccinimide ester, perfluorophenyl ester, or urethane protected acid, and the like. Other activated esters and reaction conditions are well known (See Atherton et al.).

2. Use of Photoremovable Groups During Solid-Phase Synthesis of Oligonucleotides

The formation of oligonucleotides on a solid-phase support requires the stepwise attachment of a nucleotide to a substrate-bound growing oligomer. In order to prevent unwanted polymerization of the monomeric nucleotide under the reaction conditions, protection of the 5'-hydroxyl group of the nucleotide is required. After the monomer is coupled to the end of the oligomer, the 5'-hydroxyl protecting group is removed, and another nucleotide is coupled to the chain. This cycle of coupling and deprotecting is continued for each nucleotide in the oligomer sequence. See Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, London, incorporated herein by reference for all purposes. As described above, the use of a photoremovable protecting group allows removal, via patterned irradiation, of selected portions of the substrate surface during the deprotection cycle of the solid phase synthesis. This selectively allows spatial control of the synthesis-the next nucleotide is coupled only to the irradiated areas.

Oligonucleotide synthesis generally involves coupling an activated phosphorous derivative on the 3'-hydroxyl group of a nucleotide with the 5'-hydroxyl group of an oligomer bound to a solid support. Two major chemical methods exist to perform this coupling: the phosphate-triester and phosphoramidite methods (See Gait). Protecting groups of the present invention are suitable for use in either method.

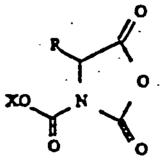
In a preferred embodiment, a photoremovable protecting group is attached to an activated nucleotide on the 5'-hydroxyl group:



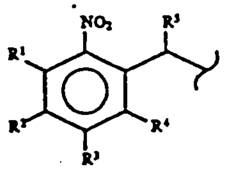
where B is the base attached to the sugar ring; R is a hydrogen atom when the sugar is deoxyribose or R is a hydroxyl group when the sugar is ribose; P represents an activated phosphorous group; and X is a photoremovable protecting group. The photoremovable protecting group, X, is preferably NV, NP, PyR, MeNV, MeNP, and the like as described above. The activated phosphorous group, P, is preferably a reactive derivative having a high coupling efficiency, such as a phosphate-triester, phosphoramidite or the like. Other activated phosphorous derivatives, as well as reaction conditions, are well known (See Gait).

E. Amino Acid N-Carboxy Anhydrides Protected With a Photoremovable Group

During Merrifield peptide synthesis, an activated ester of one amino acid is coupled with the free amino terminus of a substrate-bound oligomer. Activated esters of amino acids suitable for the solid phase synthesis include halo-formate, mixed anhydride, imidazoyl formate, acyl halide, and also includes formation of the activated ester in situ and the use of common reagents such as DCC and the like (See Atherton et al.). A preferred protected anact activated amino acid has the general formula:



where R is the side chain of the amino acid and X is a photoremovable protecting group. This compound is a urethane-protected amino acid having a photoremovable protecting group attach to the amine. A more preferred activated amino acid is formed when the photoremovable protecting group has the general formula:

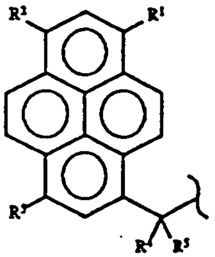


where R^1, R^2, R^3, and R^4 independently are a hydrogen atom, a lower alkyl, aryl, benzyl, halogen, hydroxyl, alkoxyl, thiol, thioether, amino, nitro, carboxyl, formate, formamido or phosphido group, or adjacent substituents (i.e., R^1-R^2, R^2-R^3, R^3-R^4) are substituted oxygen groups that together form a cyclic acetal or ketal; and R^5 is a hydrogen atom, an alkoxyl, alkyl, hydrogen, halo, aryl, or alkenyl group.

A preferred activated amino acid is formed when the photoremovable protecting group is 6-nitroveratryloxycarbonyl. That is, R^1 and R^4 are each a hydrogen atom, R^2 and R^3 are each a methoxy group, and $R^5$is a hydrogen atom. Another preferred activated amino acid is formed when the photoremovable group is

6-nitropiperonyl: R^1 and R^4 are each a hydrogen atom. R^2 and R^3 together form a methylene acetal, and R^5 is a hydrogen atom. Other protecting groups are possible. Another preferred activated ester is formed when the photoremovable group is methyl-6-nitroveratryl or methyl-6-nitropiperonyl.

Another preferred activated amino acid is formed when the photoremovable protecting group has the general formula:



where R^1, R^2, and R^3 independently are a hydrogen atom, a lower alkyl, aryl, benzyl, halogen, hydroxyl, alkoxyl, thiol, thioether, amino, nitro, carboxyl, formate, formamido, sulfanates, sulfido or phosphido group, and R^4 and R^5 independently are a hydrogen atom, an alkoxy, alkyl, halo, aryl, hydrogen, or alkenyl group. The resulting compound is a urethane-protected amino acid having a pyrenylmethyloxycarbonyl protecting group attached to the amine. A more preferred embodiment is formed when R^1 through R^5 are each a hydrogen atom.

The urethane-protected amino acids having a photoremovable protecting group of the present invention are prepared by condensation of an N-protected amino acid with an acylating agent such as an acyl halide, anhydride, chloroformate and the like (See Fuller et al., U.S. Pat. No. 4,946,942 and Fuller et al., *J. Amer. Chem. Soc.* (1990) 112:7414–7416, both herein incorporated by reference for all purposes).

Urethane-protected amino acids having photoremovable protecting groups are generally useful as reagents during solid-phase peptide synthesis, and because of the spatially selectivity possible with the photoremovable protecting group, are especially useful for the spatially addressable peptide synthesis. These amino acids are difunctional: the urethane group first serves to activate the carboxy terminus for reaction with the amine bound to the surface and, once the peptide bond is formed, the photoremovable protecting group protects the newly formed amino terminus from further reaction. These amino acids are also highly reactive to nucleophiles, such as deprotected amines on the surface of the solid support, and due to this high reactivity, the solid-phase peptide coupling times are significantly reduced, and yields are typically higher.

IV. Data Collection

A. Data Collection System

Substrates prepared in accordance with the above description are used in one embodiment to determine which of the plurality of sequences thereon bind to a receptor of interest. FIG. 11 illustrates one embodiment of a device used to detect regions of a substrate which contain flourescent markers. This device would be used, for example, to detect the presence or absence of a labeled receptor such as an antibody which has bound to a synthesized polymer on a substrate.

Light is directed at the substrate from a light source 1002 such as a laser light source of the type well known to those of skill in the art such as a model no. 2025 made by Spectra Physics. Light from the source is directed at a lens 1004 which is preferably a cylindrical lens of the type well known to those of skill in the art. The resulting output from the lens 1004 is a linear beam rather than a spot of light, resulting in the capability to detect data substantially simultaneously along a linear array of pixels rather than on a pixel-by-pixel basis. It will be understood that a cylindrical lens is used herein as an illustration of one technique for generating a linear beam of light on a surface, but that other techniques could also be utilized.

The beam from the cylindrical lens is passed through a dichroic mirror or prism (1006) and directed at the surface of the suitably prepared substrate 1008. Substrate 1008 is placed on an x-y translation stage 1009 such as a model no. PM500-8 made by Newport. Light at certain locations on the substrate will be fluoresced and transmitted along the path indicated by dashed lines back through the dichroic mirror, and focused with a suitable lens 1010 such as an f/1.4 camera lens on a linear detector 1012 via a variable f stop focusing lens 1014. Through use of a linear light beam, it becomes possible to generate data over a line of pixels (such as about 1 cm) along the substrate, rather than from individual points on the substrate. In alternative embodiments, light is directed at a 2-dimensional area of the substrate and fluoresced light detected by a 2-dimensional CCD array. Linear detection is preferred because substantially higher power densities are obtained.

Detector 1012 detects the amount of light fluoresced from the substrate as a function of position. According to one embodiment the detector is a linear CCD array of the type commonly known to those of skill in the art. The x-y translation stage, the light source, and the detector 1012 are all operably connected to a computer 1016 such as an IBM PC-AT or equivalent for control of the device and data collection from the CCD array.

In operation, the substrate is appropriately positioned by the translation stage. The light source is then illuminated, and intensity data are gathered with the computer via the detector.

FIG. 12 illustrates the architecture of the data collection system in greater detail. Operation of the system occurs under the direction of the photon counting program 1102 (photon), included herewith as Appendix B. The user inputs the scan dimensions, the number of pixels or data points in a region, and the scan speed to the counting program. Via a GPIB bus 1104 the program (in an IBM PC compatible computer, for example) interfaces with a multichannel scaler 1106 such as a Stanford Research SR 430 and an x-y stage controller 1108 such as a PM500. The signal from the light from the fluorescing substrate enters a photon counter 1110, providing output to the scaler 1106. Data are output from the scaler indicative of the number of counts in a given region. After scanning a selected area, the stage controller is activated with commands for acceleration and velocity, which in turn drives the scan stage 1112 such as a PM500-A to another region.

Data are collected in an image data file 1114 and processed in a scaling program 1116, also included in Appendix B. A scaled image is output for display on, for example, a VGA display 1118. The image is scaled based on an input of the percentage of pixels to clip and the minimum and maximum pixel levels to be viewed. The system outputs for use the min and max pixel levels in the raw data.

B. Data Analysis

The output from the data collection system is an array of data indicative of fluorescent intensity versus location on the substrate. The data are typically taken over regions substantially smaller than the area in which synthesis of a given polymer has taken place. Merely by way of example, if polymers were synthesized in squares on the substrate having dimensions of 500 microns by 500 microns, the data may be taken over regions having dimensions of 5 microns by 5 microns. In most preferred embodiments, the regions over which flourescence data are taken across the substrate are less than about ½ the area of the regions in which individual polymers are synthesized, preferably less than 1/10 the area in which a single polymer is synthesized, and most preferably less than 1/100 the area in which a single polymer is synthesized. Hence, within any area in which a given polymer has been synthesized, a large number of fluorescence data points are collected.

A plot of number of pixels versus intensity for a scan of a cell when it has been exposed to, for example, a labeled antibody will typically take the form of a bell curve, but spurious data are observed, particularly at higher intensities. Since it is desirable to use an average of fluorescent intensity over a given synthesis region in determining relative binding affinity, these spurious data will tend to undesirably skew the data.

Accordingly, in one embodiment of the invention the data are corrected for removal of these spurious data points, and an average of the data points is thereafter utilized in determining relative binding efficiency.

FIG. 13 illustrates one embodiment of a system for removal of spurious data from a set of fluorescence data such as data used in affinity screening studies. A user or the system inputs data relating to the chip location and cell corners at step 1302. From this information and the image file, the system creates a computer representation of a histogram at step 1304, the histogram (at least in the form of a computer file) plotting number of data pixels versus intensity.

For each cell, a main data analysis loop is then performed. For each cell, at step 1306, the system calculates the total intensity or number of pixels for the bandwidth centered around varying intensity levels. For example, as shown in the plot to the right of step 1306, the system calculates the number of pixels within the band of width w. The system then "moves" this bandwidth to a higher center intensity, and again calculates the number of pixels in the bandwidth. This process is repeated until the entire range of intensities has been scanned, and at step 1308 the system determines which band has the highest total number of pixels. The data within this bandwidth are used for further analysis. Assuming the bandwidth is selected to be reasonably small, this procedure will have the effect of eliminating spurious data located at the higher intensity levels. The system then repeats at step 1310 if all cells have been evaluated, or repeats for the next cell.

At step 1312 the system then integrates the data within the bandwidth for each of the selected cells, sorts the data at step 1314 using the synthesis procedure file, and displays the data to a user on, for example, a video display or a printer.

V. Representative Applications

A. Oligonucleotide Synthesis

The generality of light directed spatially addressable parallel chemical synthesis is demonstrated by application to nucleic acid synthesis.

1. Example

Light activated formation of a thymidinecytidine dimer was carried out. A three dimensional representation of a fluorescence scan showing a checkerboard pattern generated by the light-directed synthesis of a dinucleotide is shown in FIG. 8. 5'-nitroveratryl thymidine was attached to a synthesis substrate through the 3' hydroxyl group. The nitroveratryl protecting groups were removed by illumination through a 500 mm checkerboard mask. The substrate was then treated with phosphoramidite activated 2'-deoxycytidine. In order to follow the reaction fluorometrically, the deoxycytidine had been modified with an FMOC protected aminohexyl linker attached to the exocyclic amine (5'-O-dimethoxytrityl-4-N-(6-N-fluorenylmethylcarbamoyl-hexylcarboxy)-2'-deoxycytidine). After removal of the FMOC protecting group with base, the regions which contained the dinucleotide were fluorescently labelled by treatment of the substrate with 1 mM FITC in DMF for one hour.

The three-dimensional representation of the fluorescent intensity data in FIG. 14 clearly reproduces the checkerboard illumination pattern used during photolysis of the substrate. This result demonstrates that oligonucleotides as well as peptides can be synthesized by the light-directed method.

VI. Conclusion

The inventions herein provide a new approach for the simultaneous synthesis of a large number of compounds. The method can be applied whenever one has chemical building blocks that can be coupled in a solid-phase format, and when light can be used to generate a reactive group.

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. Merely by way of example, while the invention is illustrated primarily with regard to peptide and nucleotide synthesis, the invention is not so limited. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 21

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
Tyr Gly Gly Phe Leu
1               5
```

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Pro Gly Gly Phe Leu
1               5
```

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

```
Tyr Gly Ala Gly Phe
1               5
```

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
Tyr Gly Ala Phe Leu Ser
1               5
```

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

```
Tyr Gly Ala Phe Ser
1               5
```

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

```
Tyr Gly Ala Phe Leu
1               5
```

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
Tyr Gly Gly Phe Leu Ser
1               5
```

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
Tyr Gly Ala Phe
1
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:9:

```
Tyr Gly Ala Leu Ser
1               5
```

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:10:

```
Tyr Gly Gly Phe Ser
1               5
```

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:11:

```
Tyr Gly Ala Leu
1
```

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:12:

```
Tyr Gly Ala Phe Leu Phe
1               5
```

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:13:

```
Tyr Gly Ala Phe Phe
1               5
```

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:14:

```
Tyr Gly Gly Leu Ser
1               5
```

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:15:

```
Tyr Gly Gly Phe Leu
1               5
```

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:16:

```
Tyr Gly Ala Phe Ser Phe
1               5
```

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 7 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:17:

```
Tyr Gly Ala Phe Leu Ser Phe
1               5
```

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:18:

```
Tyr Gly Ala Phe Met Gln
1               5
```

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:19:

```
Tyr Gly Ala Phe Met
1               5
```

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:20:

```
Tyr Gly Ala Phe Gln
1               5
```

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 5 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

US005837832A

United States Patent [19]

Chee et al.

[11] Patent Number: 5,837,832

[45] Date of Patent: Nov. 17, 1998

[54] ARRAYS OF NUCLEIC ACID PROBES ON BIOLOGICAL CHIPS

[75] Inventors: Mark Chee, Palo Alto; Maureen T. Cronin, Los Altos; Stephen P. A. Fodor, Palo Alto; Xiaohua X. Huang; Earl A. Hubbell, both of Mt. View; Robert J. Lipshutz; Peter E. Lobban, both of Palo Alto; MacDonald S. Morris, San Jose; Edward L. Sheldon, Menlo Park, all of Calif.

[73] Assignee: Affymetrix, Inc., Santa Clara, Calif.

[21] Appl. No.: 441,887

[22] Filed: May 16, 1995

Related U.S. Application Data

[63] Continuation of Ser. No. 143,312, Oct. 26, 1993, abandoned, which is a continuation-in-part of Ser. No. 82,937, Jun. 25, 1993, abandoned.

[51] Int. Cl.^{6} C12Q 1/68

[52] U.S. Cl. 536/22.1; 435/6; 435/91.1; 436/501; 536/23.1; 536/24.1; 536/24.3; 536/24.31; 536/24.32; 536/24.33; 536/25.3; 422/68.1; 935/77; 935/78; 935/88

[58] Field of Search 435/6, 91.1, 810; 436/501; 536/22.1, 23.1, 24.1, 24.3–24.33, 25, 3; 935/77, 78, 88; 422/68.1

[56] References Cited

U.S. PATENT DOCUMENTS

4,656,127	4/1987	Mundy	435/6
4,683,195	7/1987	Mallis et al.	435/6
5,002,867	3/1991	Macevicz	435/6
5,143,854	9/1992	Pirrung et al.	436/518
5,202,231	4/1993	Drmanac et al.	435/6
5,273,632	12/1993	Stockham et al.	204/180.1
5,527,681	6/1996	Homes	435/6

FOREIGN PATENT DOCUMENTS

WO 89/10977	11/1989	WIPO	C12Q 1/68
WO 89/11548	11/1989	WIPO	C12Q 1/68
WO 90/00626	1/1990	WIPO	C12Q 1/68
WO 90/03382	4/1990	WIPO	C07H 21/00
WO 92/10092	6/1992	WIPO	A01N 1/02
WO 92/10588	6/1992	WIPO	C12Q 1/68
WO 93/10588	6/1992	WIPO	C12Q 1/68
WO 93/17126	9/1993	WIPO	C12Q 1/68

OTHER PUBLICATIONS

Maram et al. (1980) Methods in Enzymology, vol. 65, pp. 449–559.

Sambrook et al., Molecular Cloning, Cold Spring Harbor Laboratory Press, 1989, pp. 1145–1147

Stratagene 1988 Catalog, pp. 39.

Elder, J.K., "Analysis of DNA oligonucleotide hybridization data by maximum entropy," *Maximum Entropy and Bayesian Methods*, pp. 1–10, Paris (1992).

Lipshutz, Robert J., "Likelihood DNA sequencing by hybridization," *J. of Biomolecular Structure & Dynamics* 11:637–653 (1993).

Ying Luo et al., "Cellular protein modulates effects of human immunodeficiency virus type 1 rev," *J. of Virology* 68:3850–3856 (1994).

Querat et al., "Nucleotide sequence analysis of SA–OMVV, a Visna–related ovine lentivirus: phylogenetic history of lentiviruses," *Virology* 175:434–447 (1990).

Ratner et al., "Complete nucleotide sequence of the AIDS virus, HTLV-III," *Nature* 313:277–284 (1985).

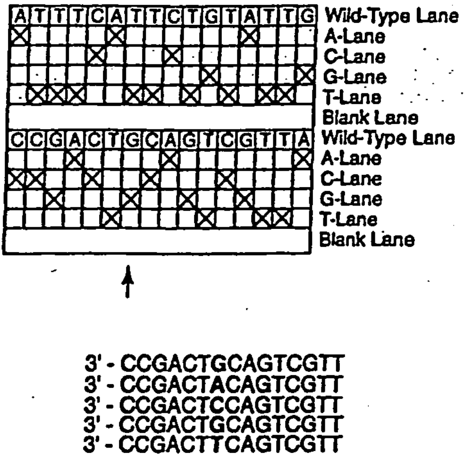
(List continued on next page.)

Primary Examiner—Ardin H. Marschel
Attorney, Agent, or Firm—Townsend & Townsend & Crew

[57] ABSTRACT

DNA chips containing arrays of oligonucleotide probes can be used to determine whether a target nucleic acid has a nucleotide sequence identical to or different from a specific reference sequence. The array of probes comprises probes exactly complementary to the reference sequence, as well as probes that differ by one or more bases from the exactly complementary probes.

18 Claims, 40 Drawing Sheets



2

OTHER PUBLICATIONS

Wain–Hobson et al., "Nucleotide sequence of the AIDS virus, LAV," *Cell* 40:9–17 (1985).

Southern et al., "Analyzing and Comparing Nucleic Acid Sequences by Hybridization to Arrays of Oligonucleotides: Evaluation Using Experimental Models," *Genomics* (1992) 13:1008–1017.

Sanger et al., "DNA sequencing with chain–terminating inhibitors," *Proc. Natl. Acad. Sci. USA* (1977) 74:5463–5467.

M. Cronin et al., Hybridization to Arrays of Oligonucleotides, Poster Presentation: Nucleic Acids In Medical Applications Conference sponsored by AACC, Jan. 1993, published in conference syllabus, Cancun, Mexico.

M.S. Chee et al., Towards Sequencing Mitochondrial DNA Polymorphisms by Hybridization to a Custom Oligonucleotide Probe Array, American Society of Human Genetics 43rd Annual Meeting, Oct. 5–9, 1993, New Orleans, LA.

M.S. Chee et al., Genetic Analysis by Hybridization to Sequence–Specific DNA Arrays, Genome Sequencing and Analysis Conference V, Oct. 23–27, 1993, Hilton Head, SC.

P.E. Lobban et al., DNA Chips for Genetic Analysis, Genome Sequencing and Analysis Conference V, Oct. 23–27, 1993, Hilton Head, SC.

R. Lipshutz et al., Oligonucleotide Arrays for Hybridization Analysis, Genome Sequencing and Analysis Conference V, Oct. 23–27, 1993, Hilton Head, SC.

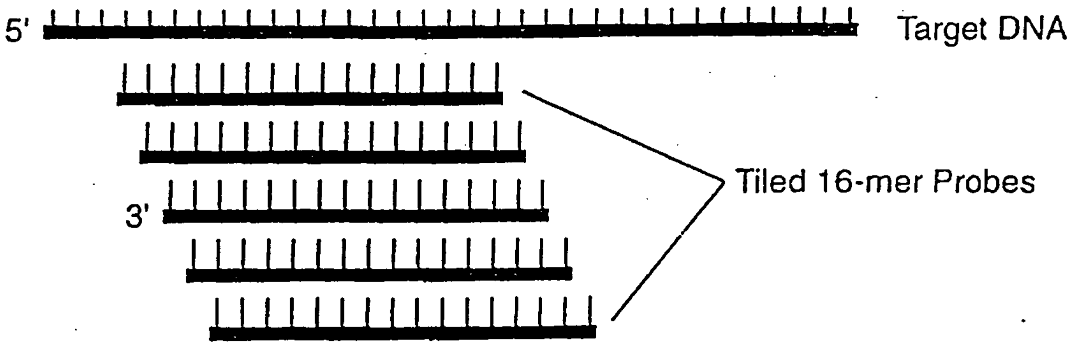
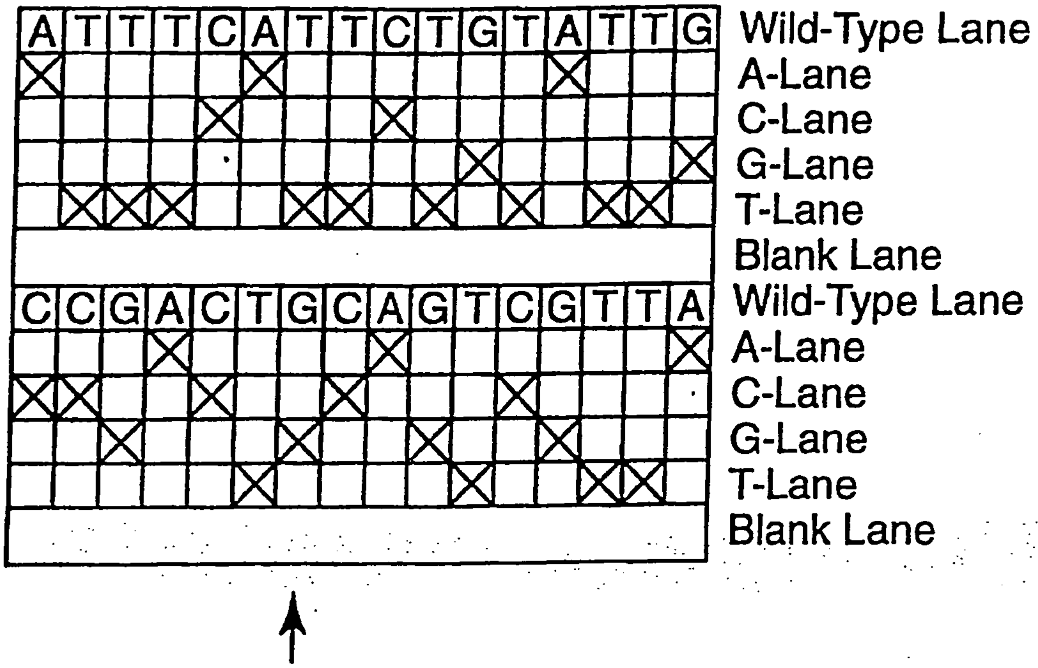


Fig. 1



A	T	T	T	C	A	T	T	C	T	G	T	A	T	T	G	Wild-Type Lane
☒					☒							☒				A-Lane
				☒				☒								C-Lane
										☒					☒	G-Lane
	☒	☒	☒			☒	☒		☒		☒		☒	☒		T-Lane
																Blank Lane
C	C	G	A	C	T	G	C	A	G	T	C	G	T	T	A	Wild-Type Lane
			☒					☒							☒	A-Lane
☒	☒			☒			☒				☒					C-Lane
		☒				☒			☒			☒				G-Lane
					☒					☒			☒	☒		T-Lane
																Blank Lane

3' - CCGACT**G**CAGTCGTT
3' - CCGACT**A**CAGTCGTT
3' - CCGACT**C**CAGTCGTT
3' - CCGACT**G**CAGTCGTT
3' - CCGACT**T**CAGTCGTT

Fig. 2

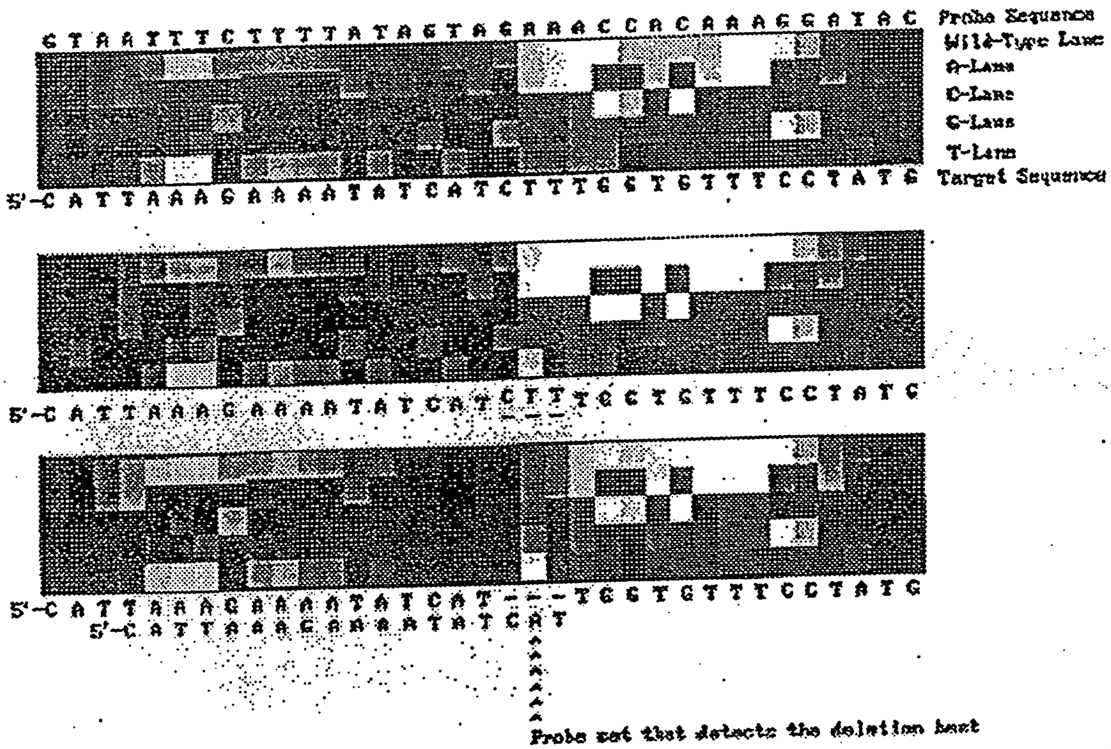

G T A A T T T C T T T T A T A G T A G A A A C C A C A A A G G A T A C
Probe Sequence
Wild-Type Lane
A-Lane
C-Lane
G-Lane
T-Lane
Target Sequence
5'-C A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G
5'-C A T T A A A G A A A A T A T C A T C T T T G C T G T T T C C T A T G
5'-C A T T A A A G A A A A T A T C A T - - - T G G T G T T T C C T A T G
5'-C A T T A A A G A A A A T A T C A T
Probe set that detects the deletion best

FIG. 3

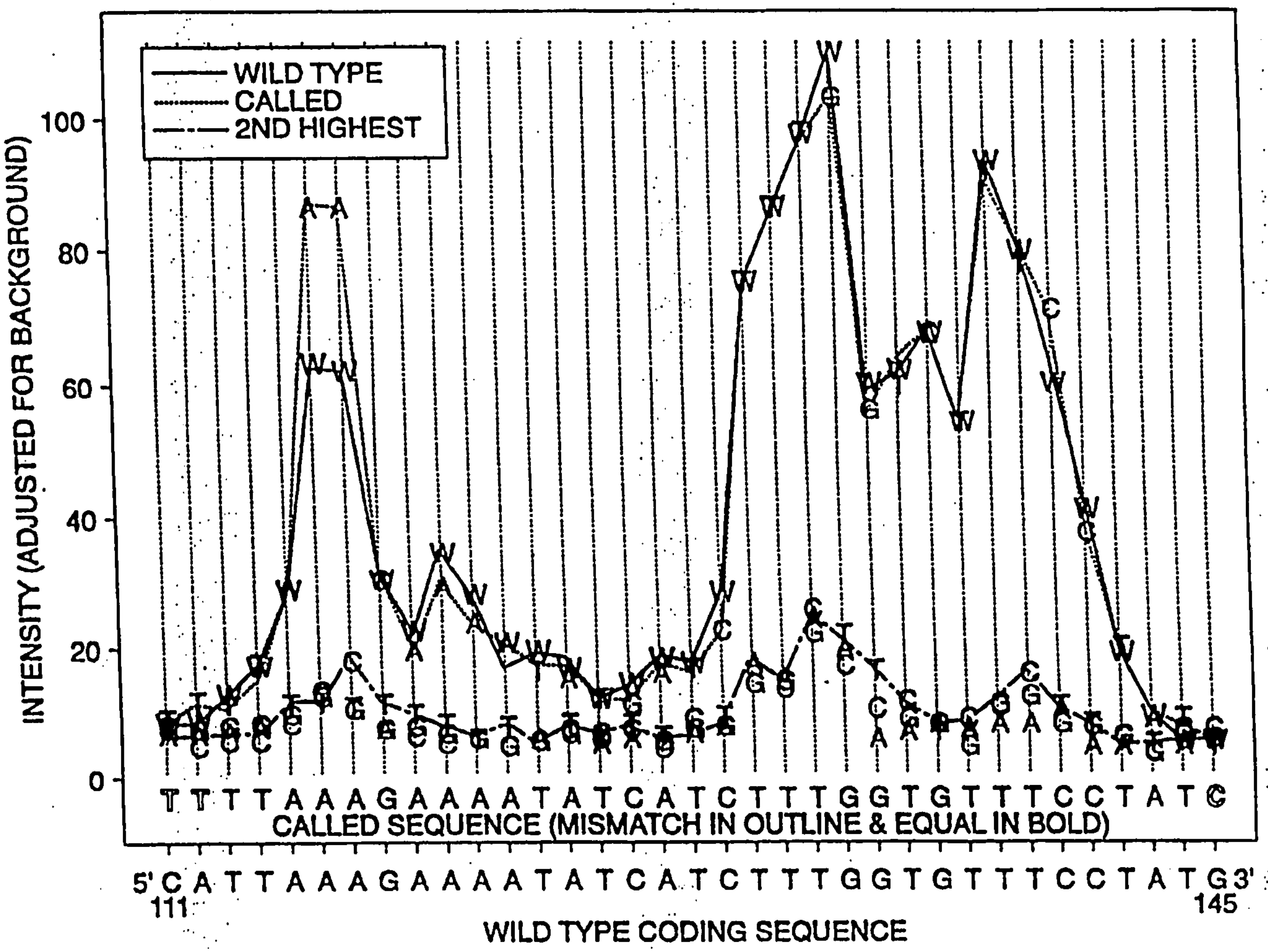

WILD TYPE
CALLED
2ND HIGHEST
INTENSITY (ADJUSTED FOR BACKGROUND)
0
20
40
60
80
100
T T T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T C
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5' C A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G 3'
111
145
WILD TYPE CODING SEQUENCE

Fig. 4A

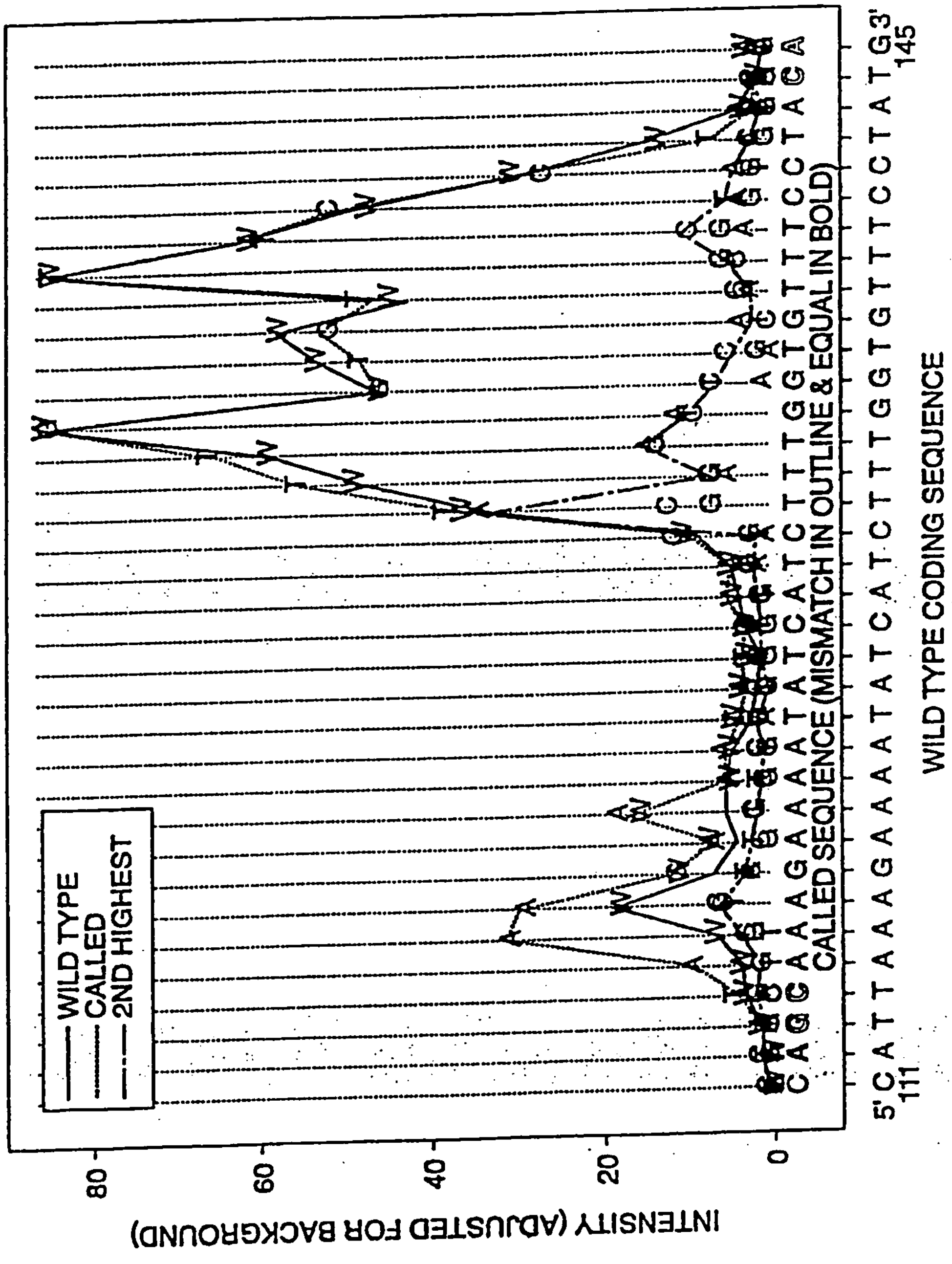

WILD TYPE
CALLED
2ND HIGHEST
INTENSITY (ADJUSTED FOR BACKGROUND)
80
60
40
20
0
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5'C A T T A A A G A A A T A T C A T C T T T G G T G T T T C C C T A T G3'
111
145
WILD TYPE CODING SEQUENCE

Fig. 4B

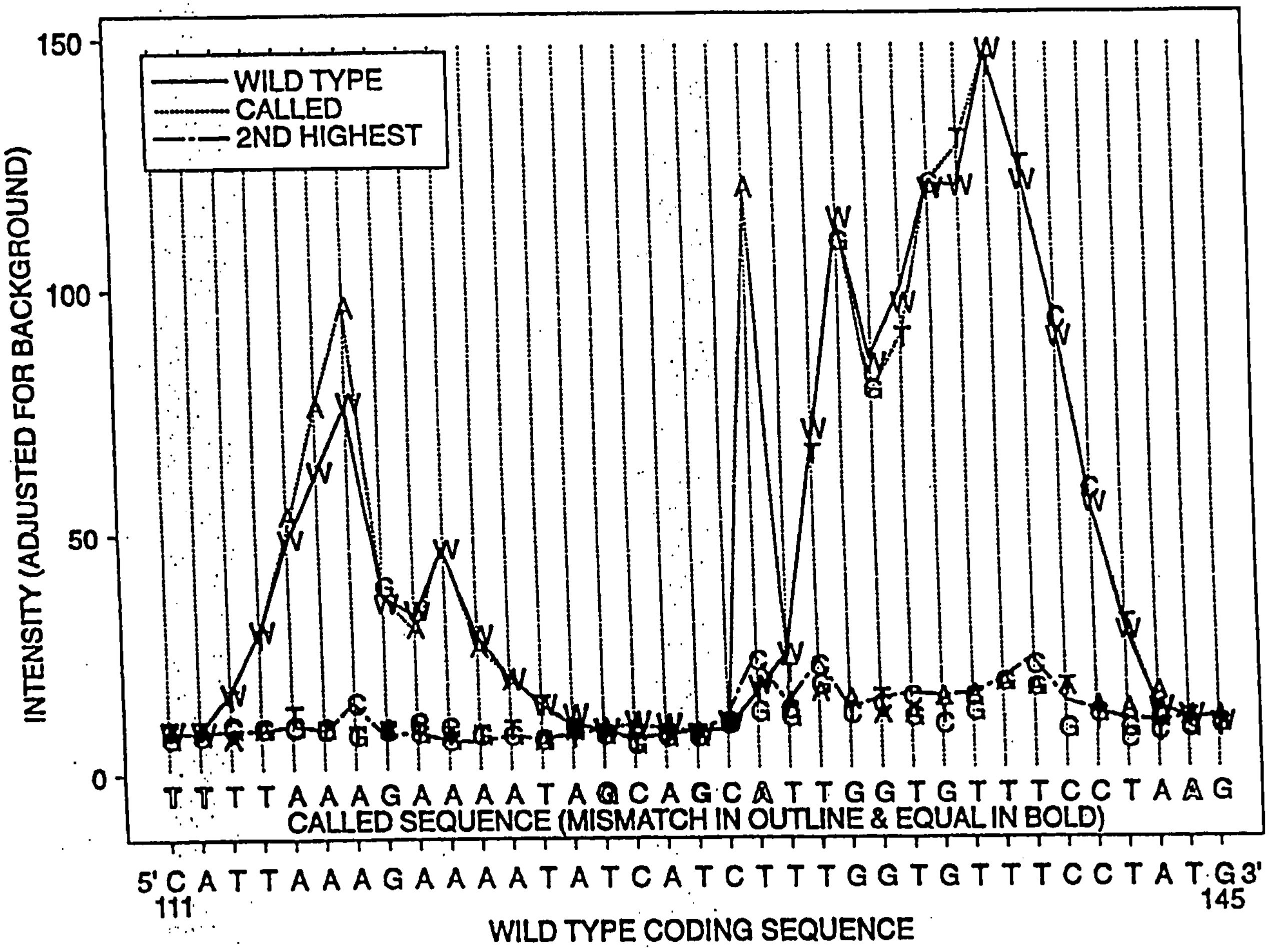

WILD TYPE
CALLED
2ND HIGHEST
150
100
50
0
INTENSITY (ADJUSTED FOR BACKGROUND)
T T T T A A A G A A A A T A G C A G C A T T G G T G T T T C C T A A G
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5'C A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G3'
111
145
WILD TYPE CODING SEQUENCE

Fig. 4C

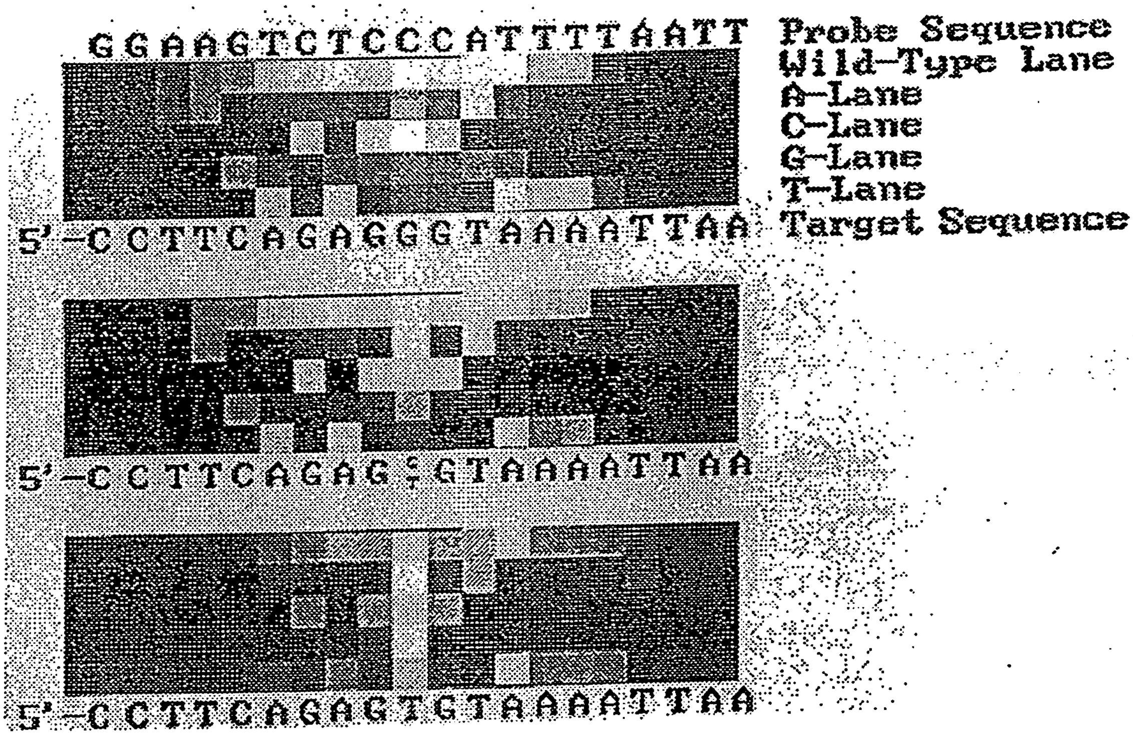

GGAAGTCTCCCATTTTAATT
Probe Sequence
Wild-Type Lane
A-Lane
C-Lane
G-Lane
T-Lane
5'-CCTTCAGAGGGTAAAATTAA
Target Sequence
5'-CCTTCAGAG C/T GTAAAATTAA
5'-CCTTCAGAGTGTAAAATTAA

FIG. 5

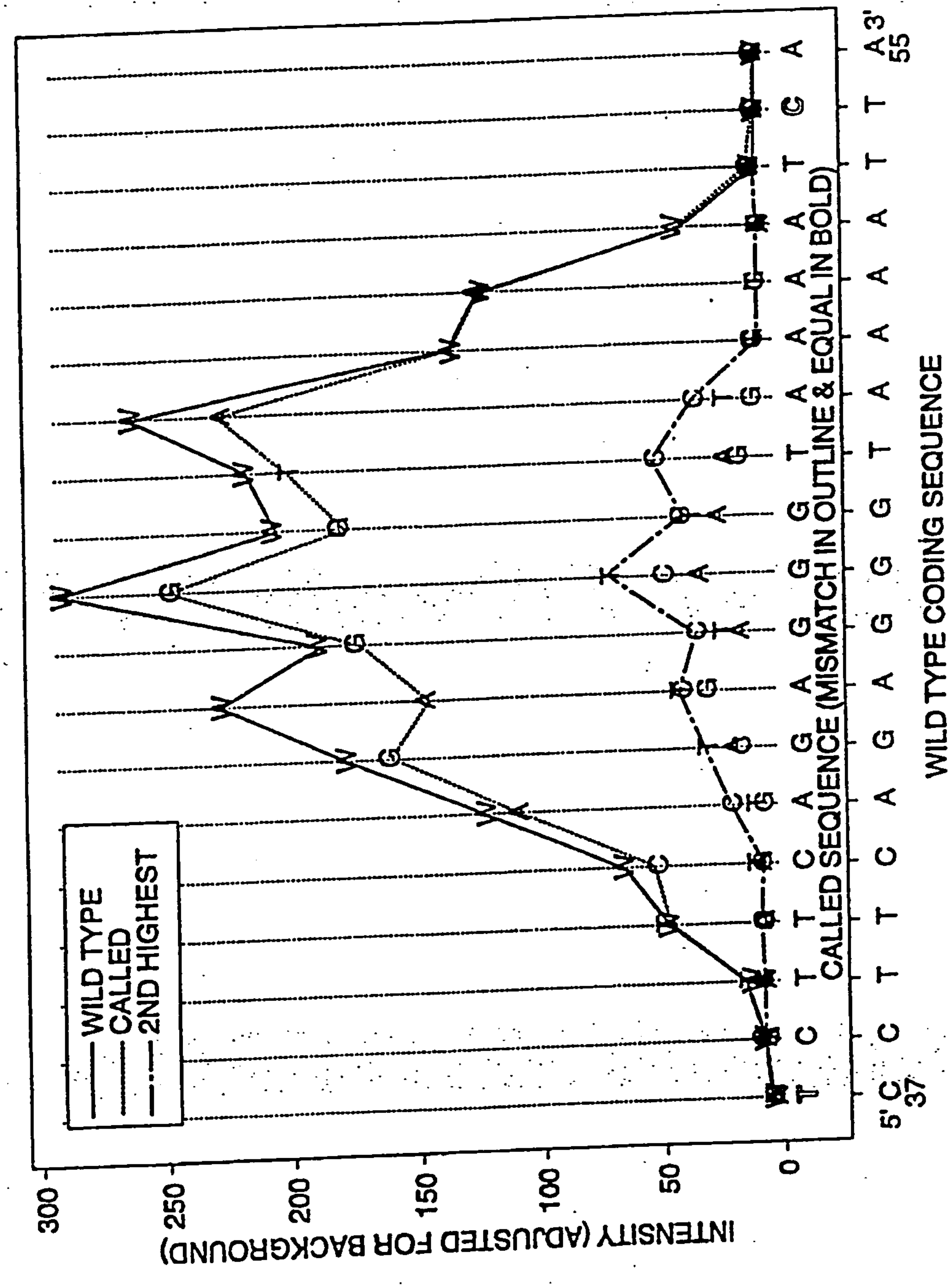

INTENSITY (ADJUSTED FOR BACKGROUND)
0
50
100
150
200
250
300
WILD TYPE
CALLED
2ND HIGHEST
T C T T C A G A G G G T G A A A A T G A
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5' C C T T C A G A G G G T A A A A T T A 3'
37
55
WILD TYPE CODING SEQUENCE

Fig. 6A

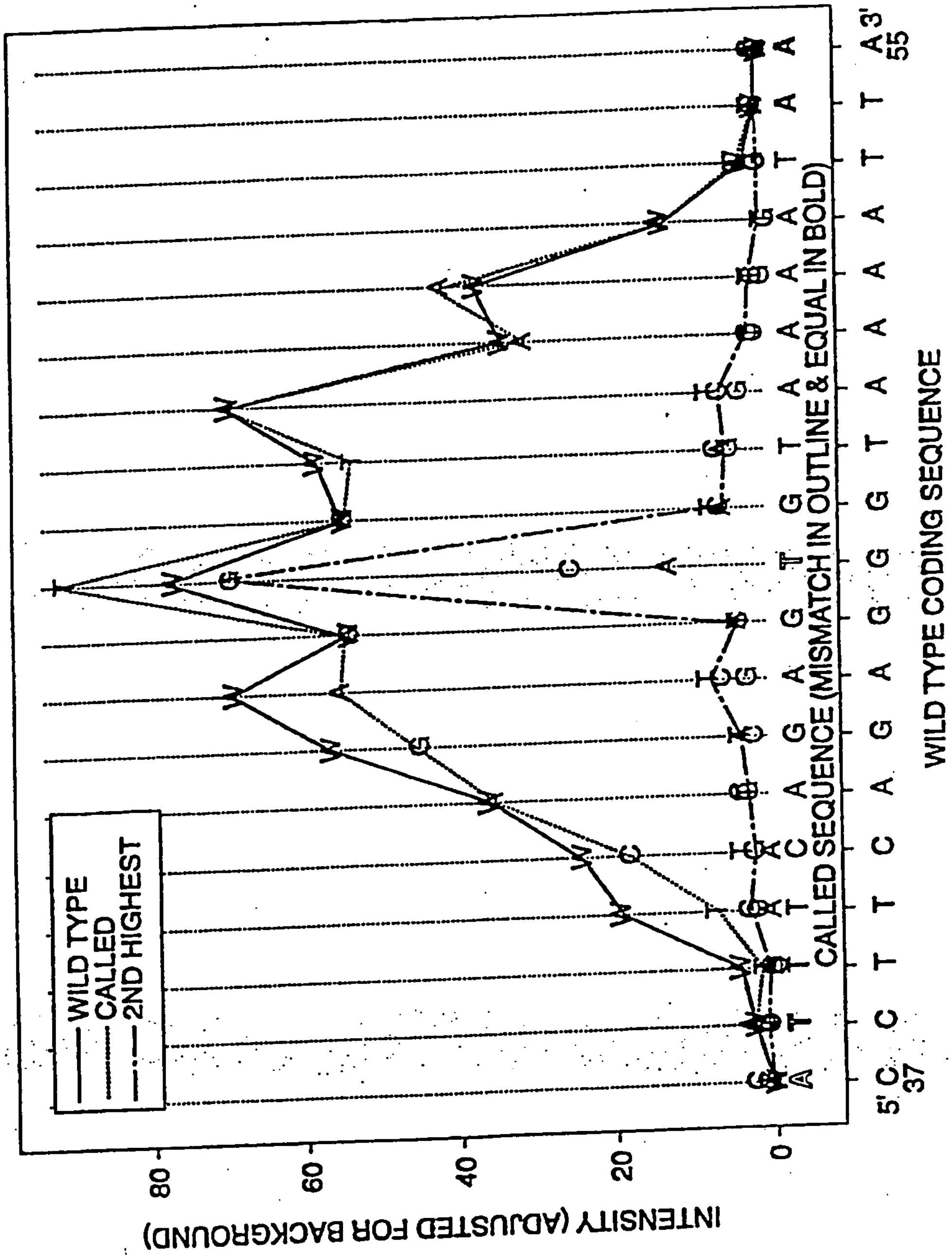

WILD TYPE
CALLED
2ND HIGHEST
INTENSITY (ADJUSTED FOR BACKGROUND)
0
20
40
60
80
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
WILD TYPE CODING SEQUENCE
5' C
37
C
T
T
C
A
G
A
G
G
G
T
A
A
A
A
T
T
A 3'
55

Fig. 6B

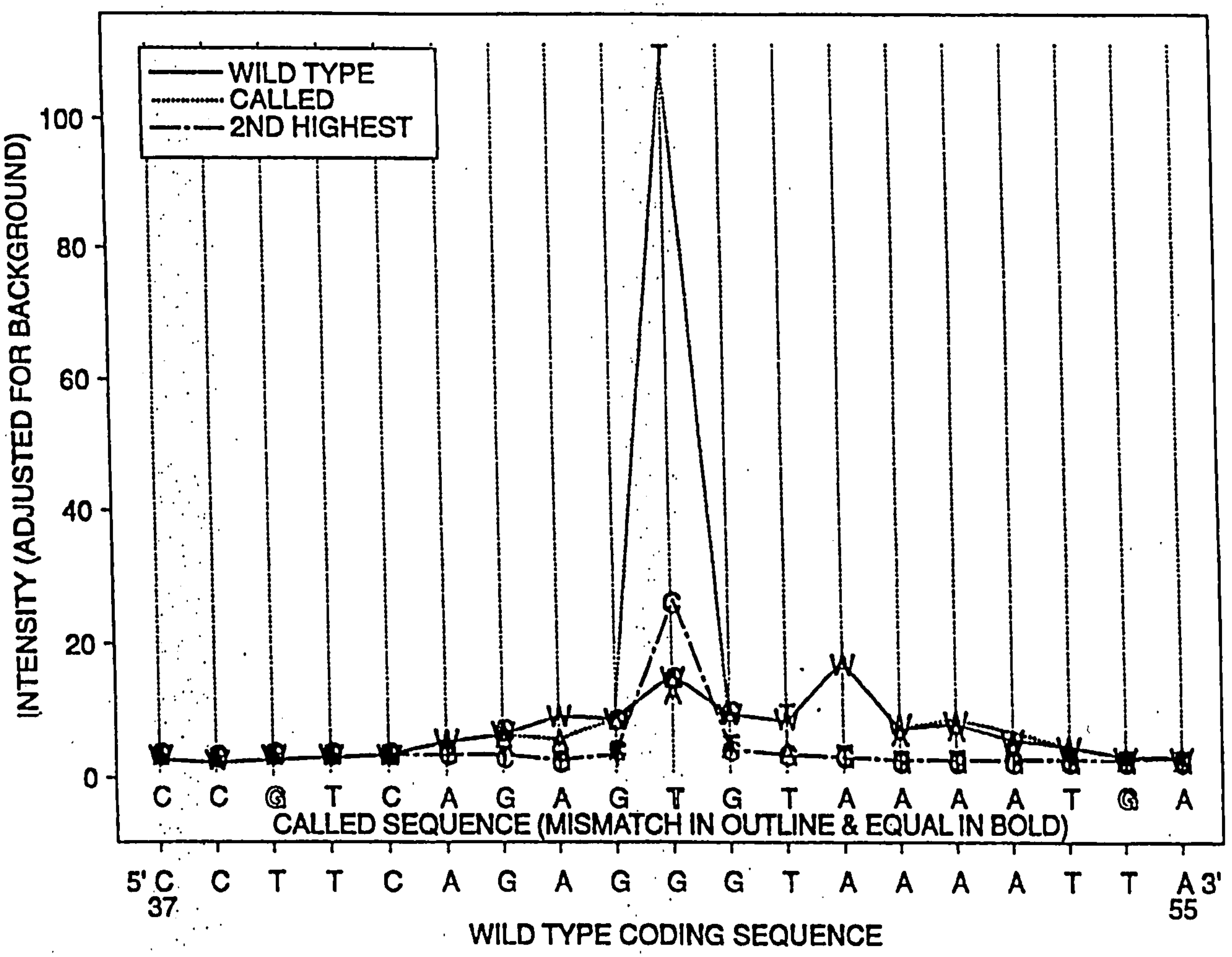

WILD TYPE
CALLED
2ND HIGHEST
INTENSITY (ADJUSTED FOR BACKGROUND)
100
80
60
40
20
0
C C G T C A G A G T G T A A A A T G A
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5' C C T T C A G A G G G T A A A A T T A 3'
37
55
WILD TYPE CODING SEQUENCE

Fig. 6C

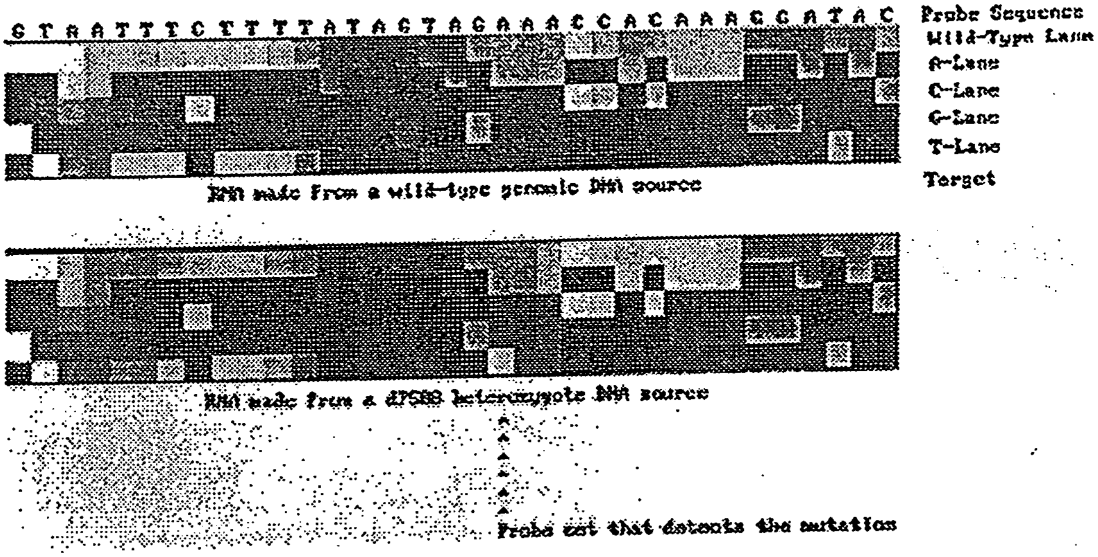

G T A A T T T C T T T T A T A G T A G A A A C C A C A A A G G A T A C
Probe Sequence
Wild-Type Lane
A-Lane
C-Lane
G-Lane
T-Lane
Target
RNA made from a wild-type genomic DNA source
RNA made from a dF508 heterozygote DNA source
Probe set that detects the mutation

FIG. 7

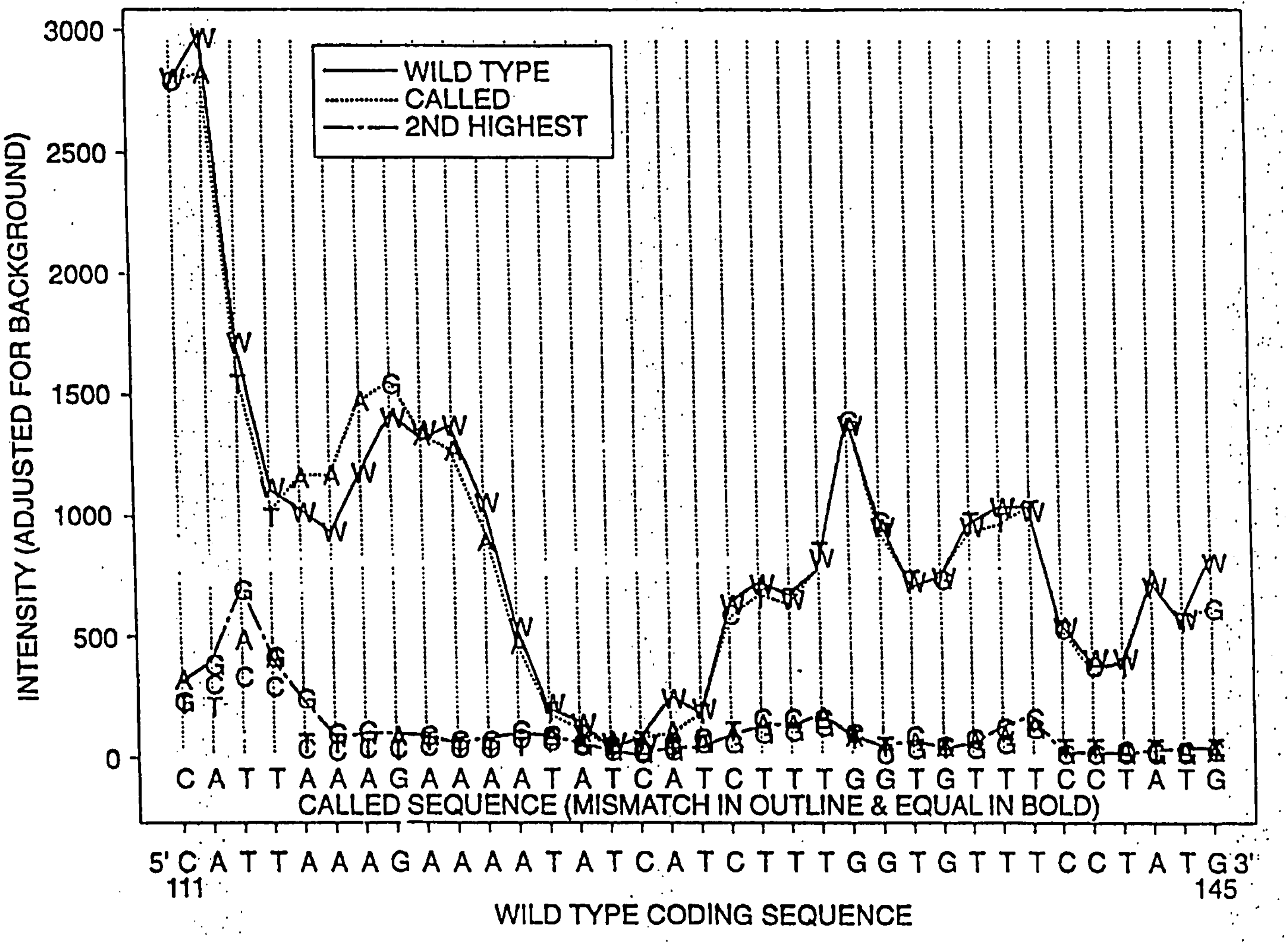

WILD TYPE
CALLED
2ND HIGHEST
INTENSITY (ADJUSTED FOR BACKGROUND)
3000
2500
2000
1500
1000
500
0
C A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5' C A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G 3'
111
145
WILD TYPE CODING SEQUENCE

Fig. 8A

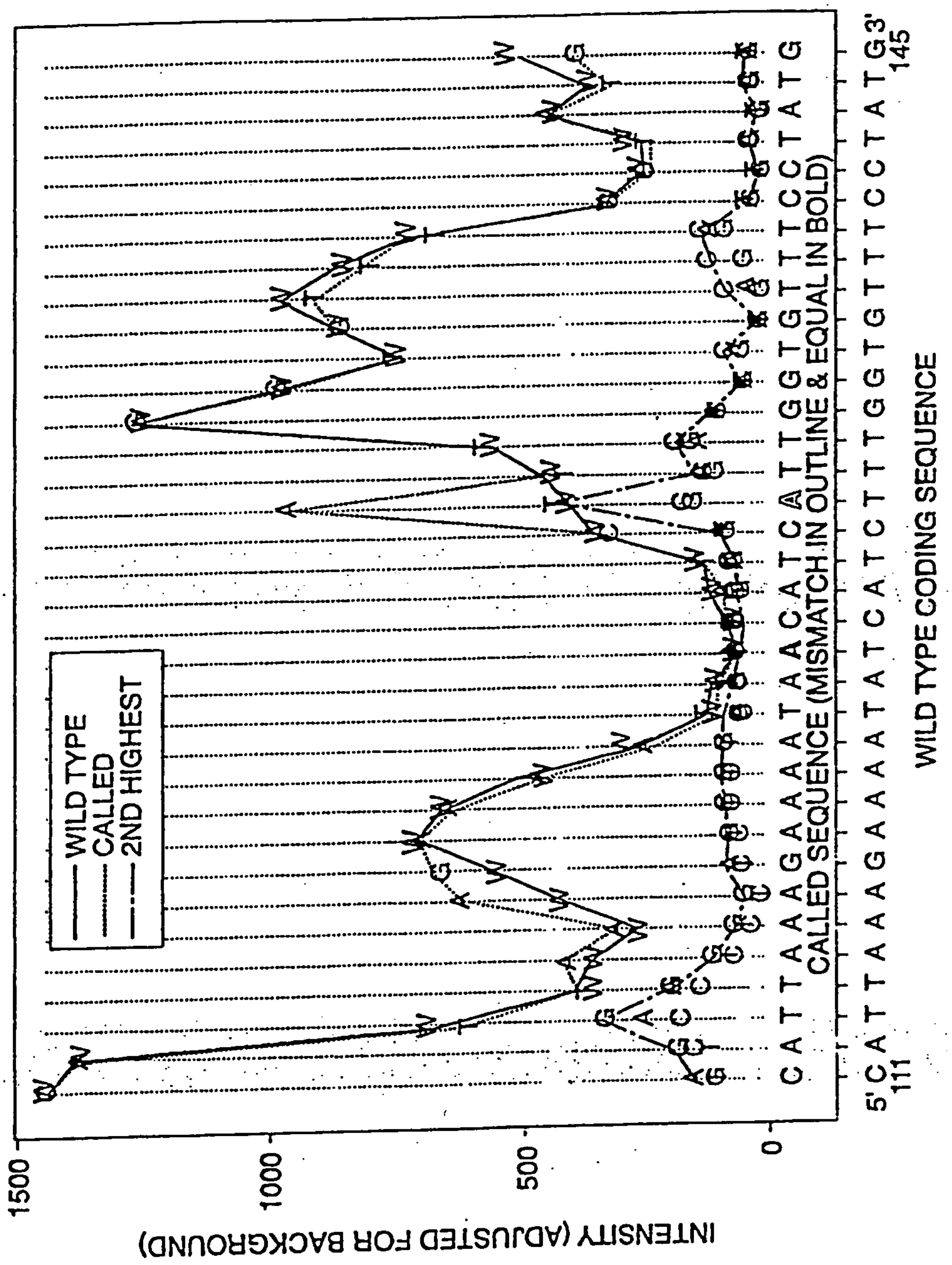

WILD TYPE
CALLED
2ND HIGHEST
1500
1000
500
0
INTENSITY (ADJUSTED FOR BACKGROUND)
C A T T A A A G A A A A T A A C A T C A T T G G T G T T T C C T A T G
CALLED SEQUENCE (MISMATCH IN OUTLINE & EQUAL IN BOLD)
5'C A T T A A A G A A A A T A T C A T C T T T G G T G T T T C C T A T G3'
111
145
WILD TYPE CODING SEQUENCE

Fig. 8B

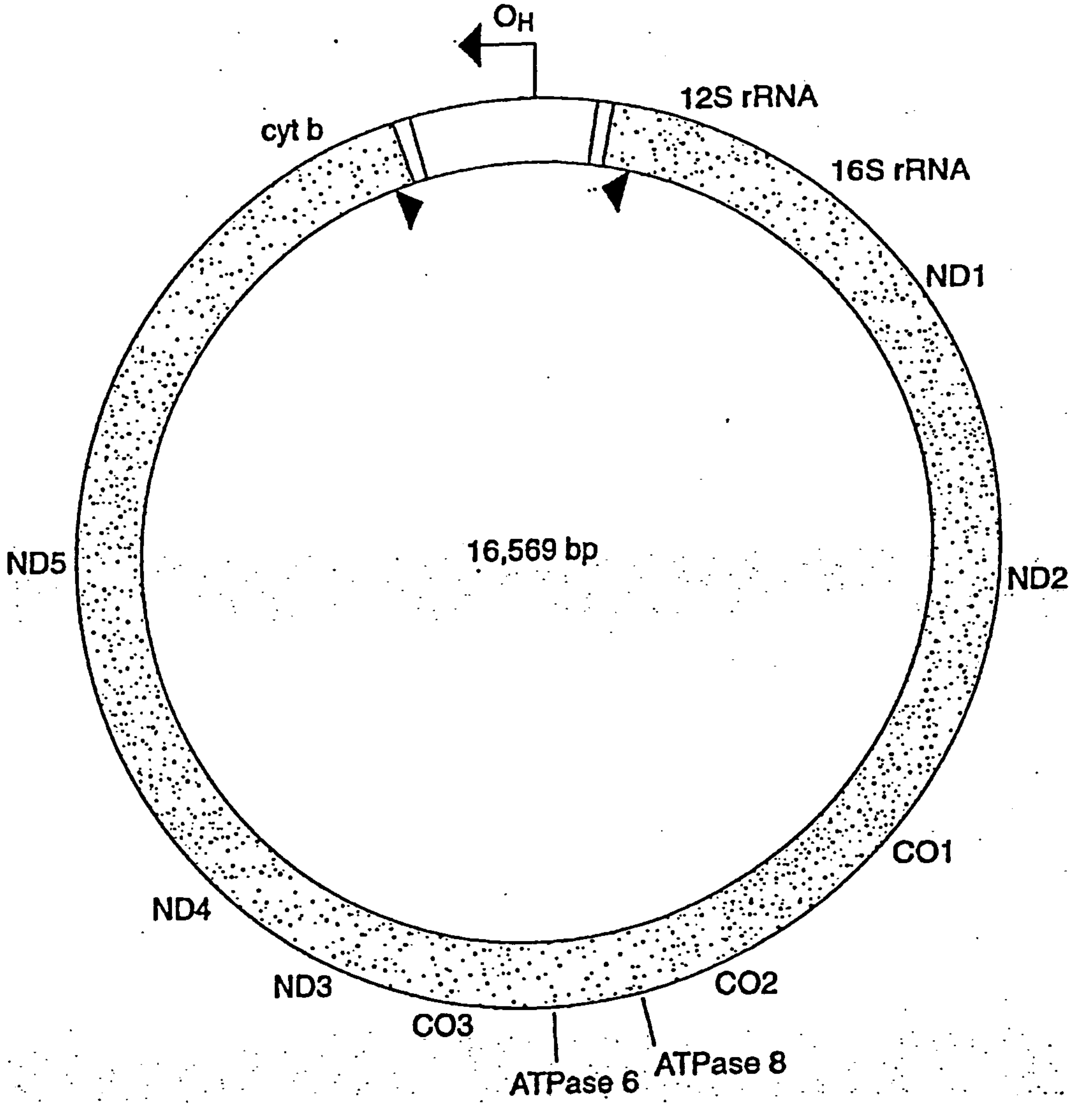
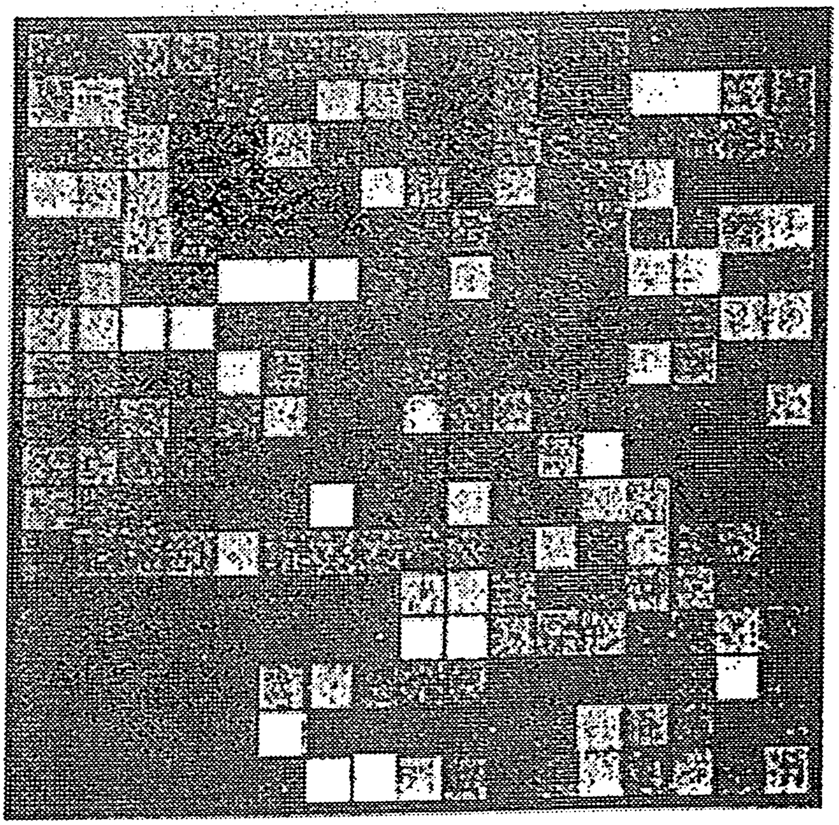

O_H
12S rRNA
16S rRNA
ND1
ND2
CO1
CO2
ATPase 8
ATPase 6
CO3
ND3
ND4
ND5
cyt b
16,569 bp

Fig. 9



mt4

FIG. 10

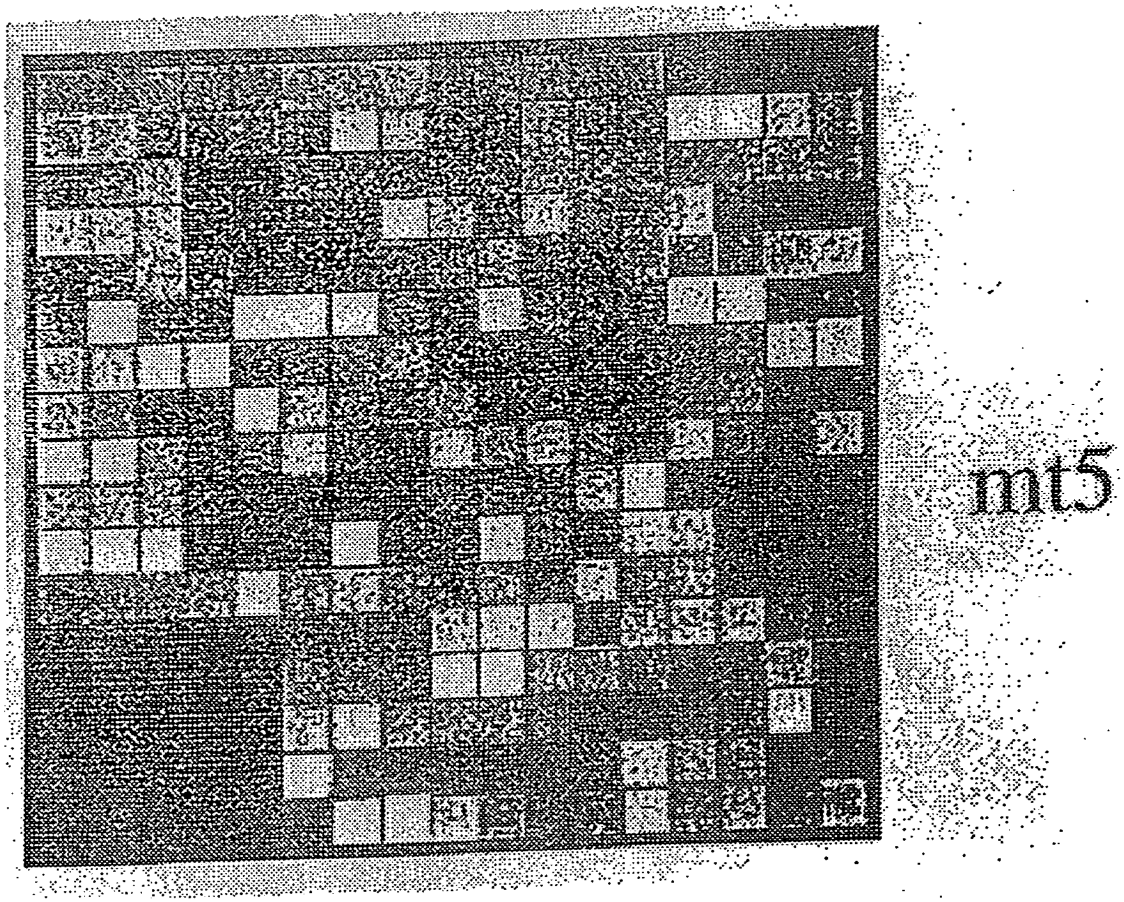


FIG. 11

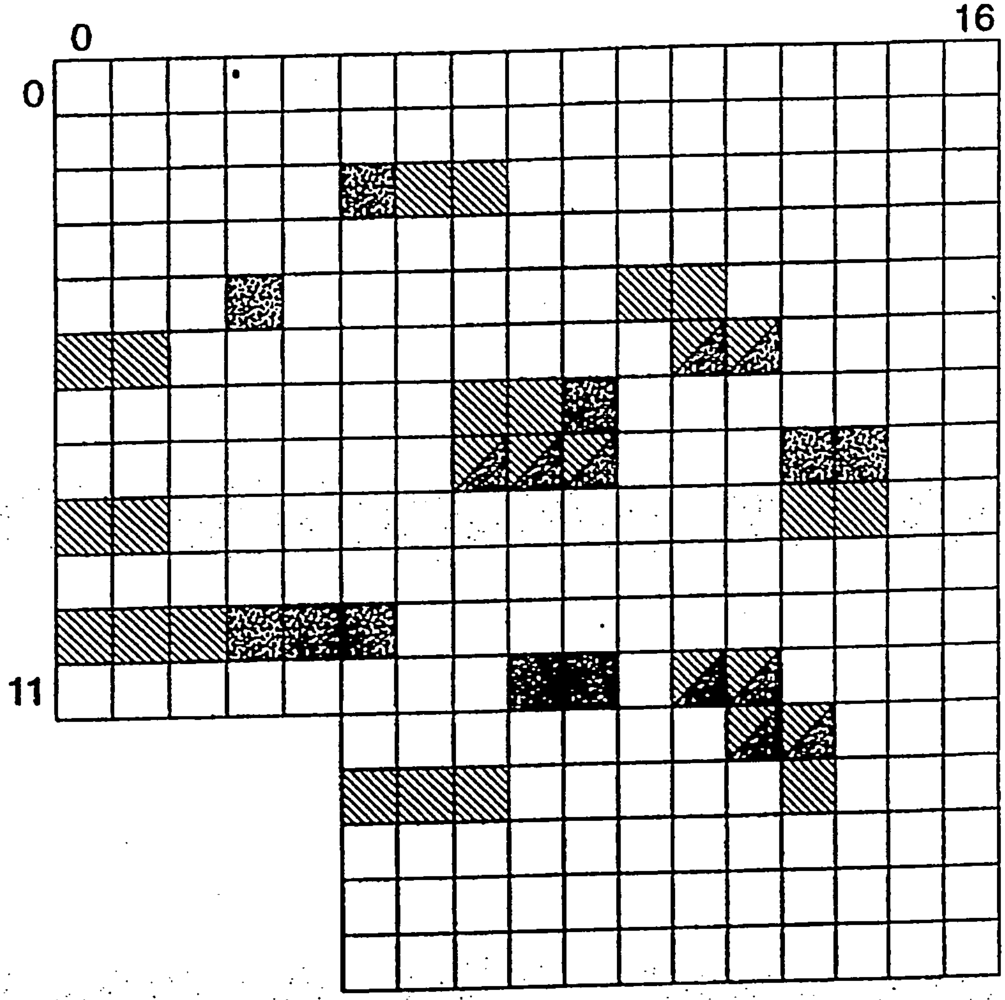


Fig. 12

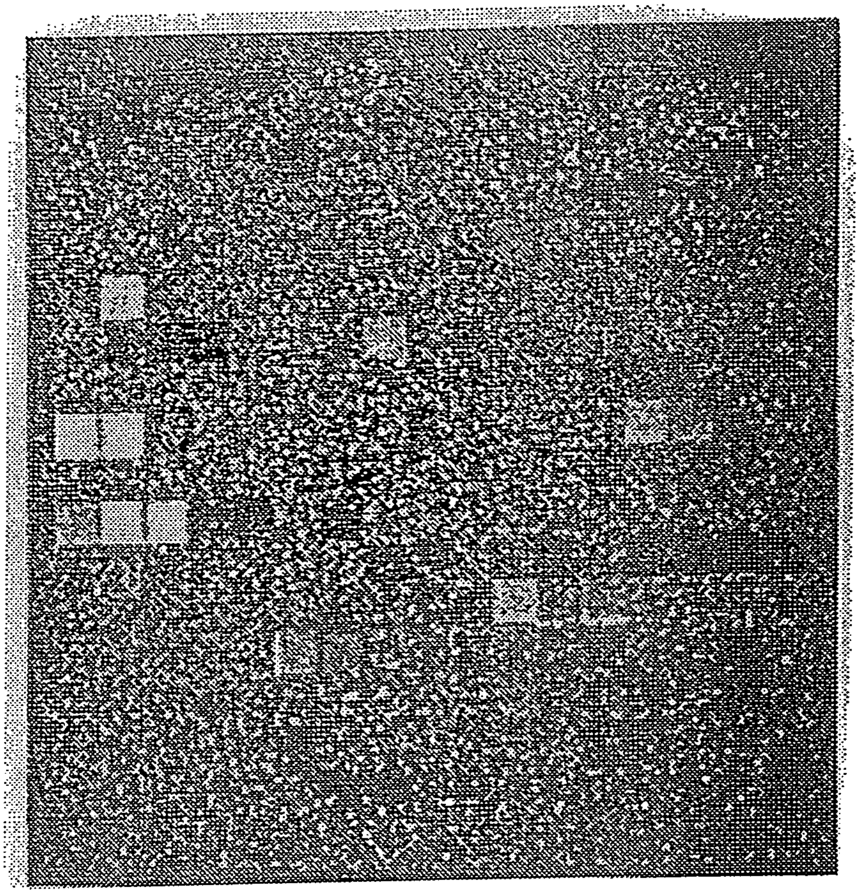
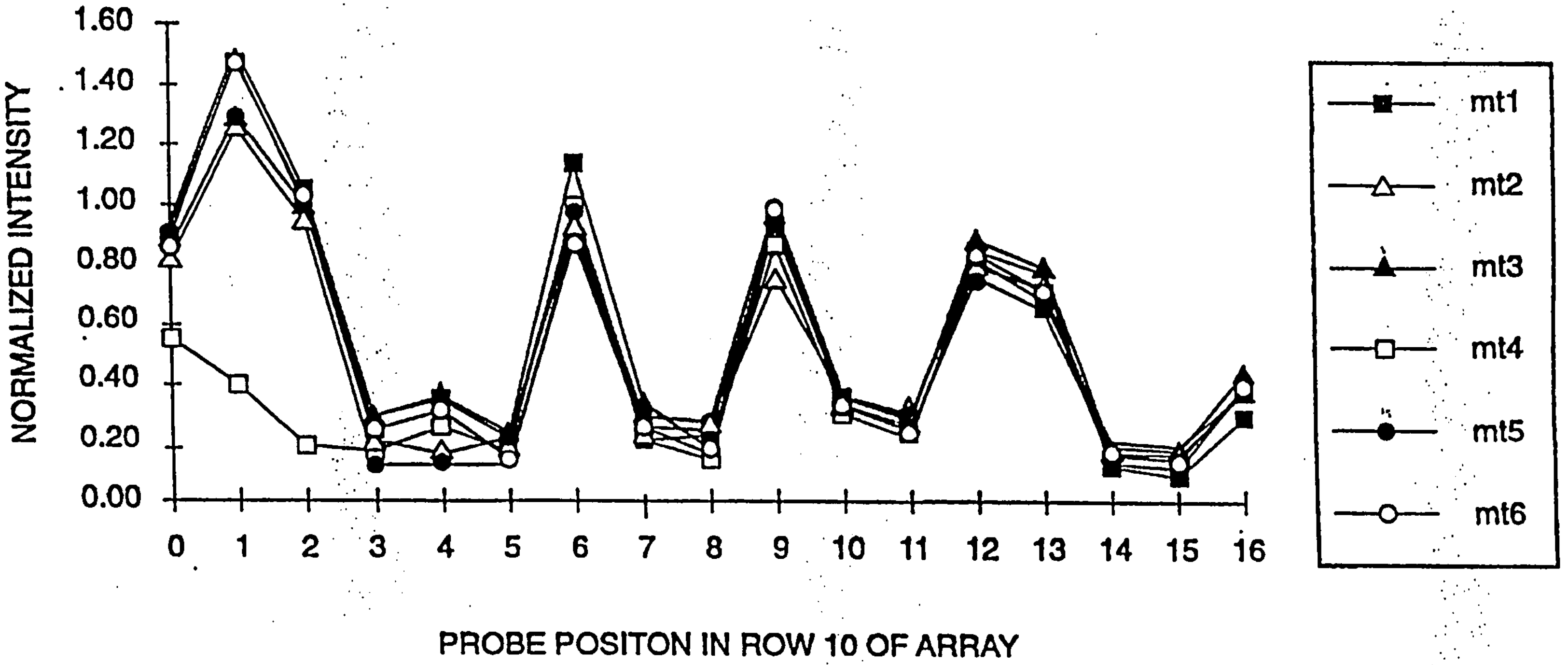
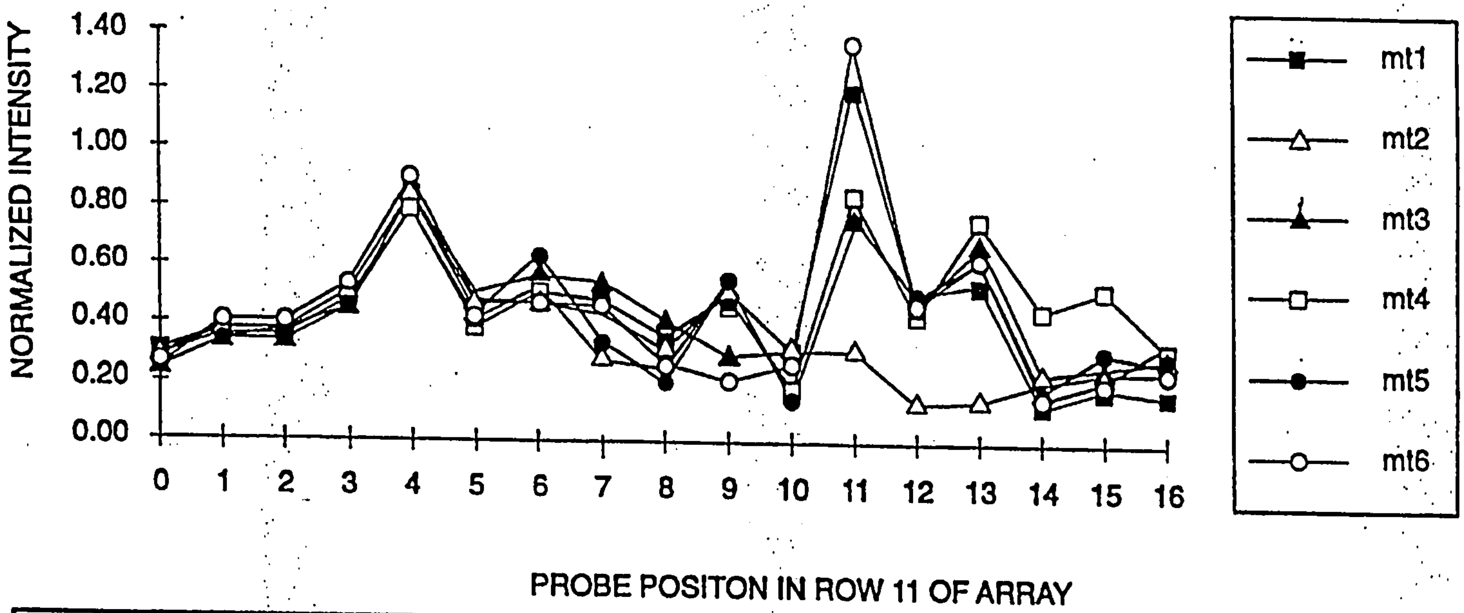


FIG. 13



PROBE POSITON	0	1	2	3	4	5
PROBE LENGTH	13	13	12	12	12	12
SAMPLE (mt1 → 6)	4	4	4	2, 5	2, 5	2, 5
MISMATCH POSITION FROM 3' OF PROBE	12	5	3	12	7	2
BASE CHANGE	t → a	t → a	t → a	t → c	t → c	t → c

Fig. 14A



PROBE POSITON	6	7	8	9	10	11	12	12
PROBE LENGTH	13	12	12	13	14	13	12	12
SAMPLE (mt1 → 6)	2	2, 5	2, 5, 6	3, 6	3, 4, 5	2, 4, 5	2	2
MISMATCH POSITION FROM 3' OF PROBE	13	9, 10	3, 4, 11	11, 5	4, 11, DOUBLE	11, 3, DOUBLE	6	3
BASE CHANGE	c → t	c → t	c → t t → c	t → c	t → c DOUBLE	g → a t → c DOUBLE	g → a	g → a

Fig. 14B

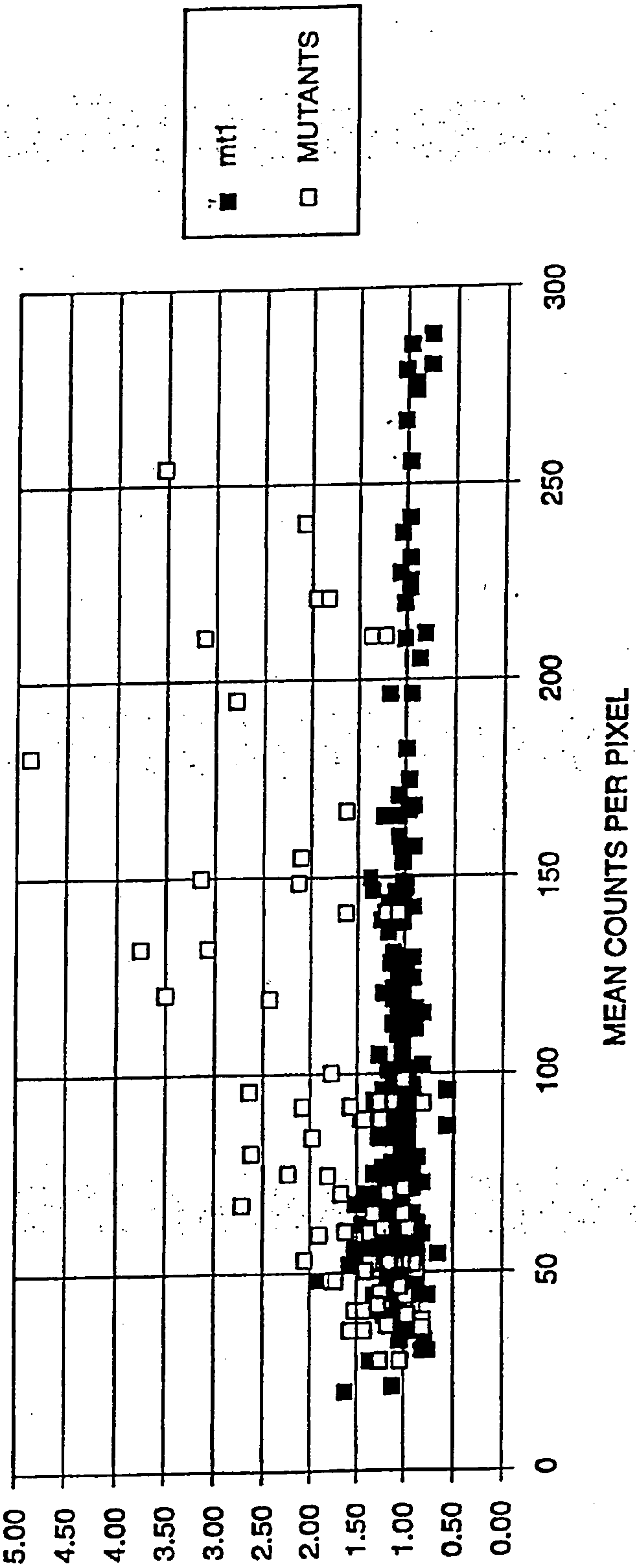

■ mt1
□ MUTANTS
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0
50
100
150
200
250
300
MEAN COUNTS PER PIXEL

Fig. 15

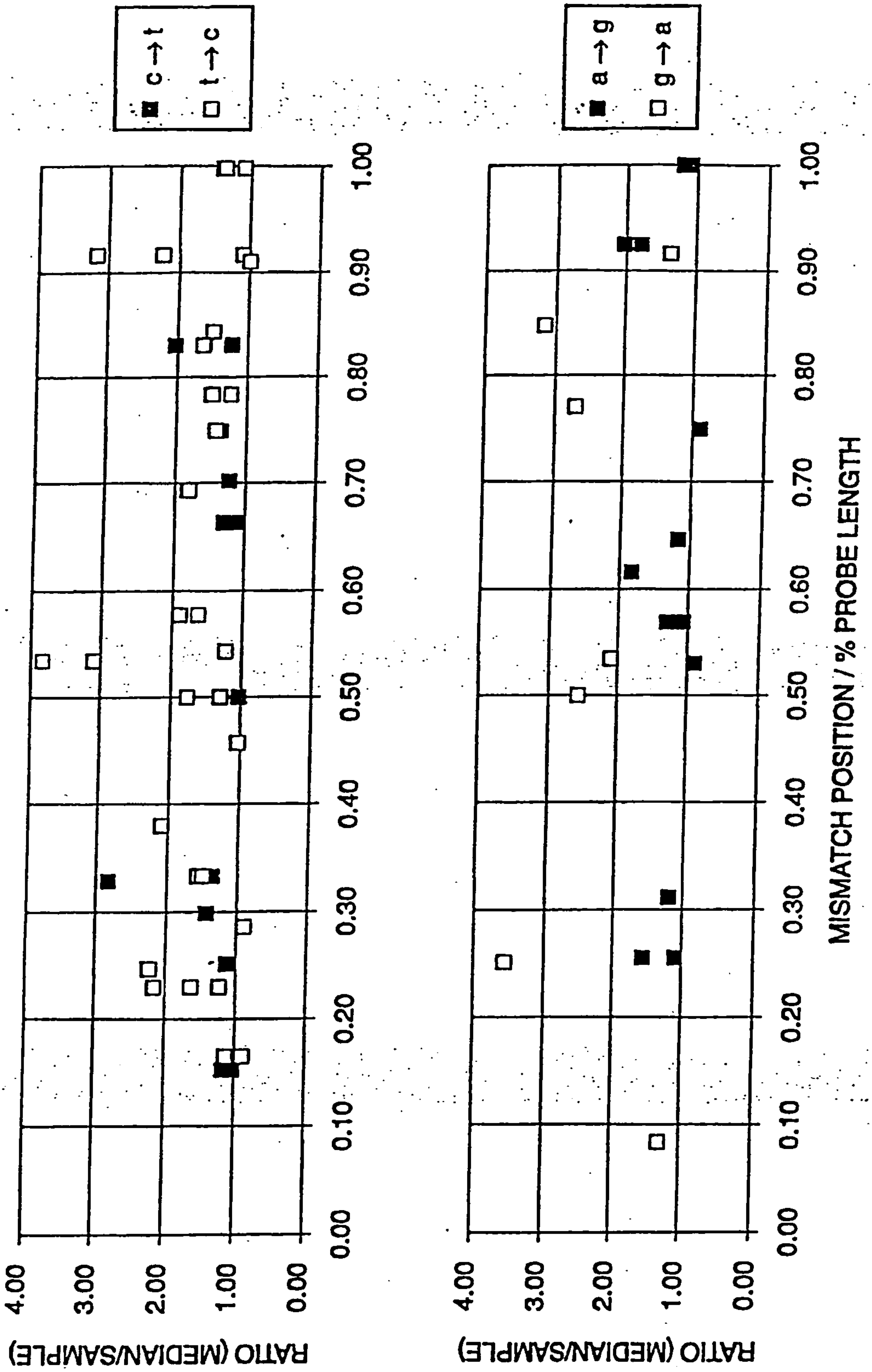

■ c → t
□ t → c
■ a → g
□ g → a
RATIO (MEDIAN/SAMPLE)
4.00
3.00
2.00
1.00
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
MISMATCH POSITION / % PROBE LENGTH

Fig. 16

```
1                                                 50
XaacaaacctacccacccttaacagtacatagtacataaagccatttacX
cgtacatagcacattacagtcaaatcccttctcgtccccatggatgaccc
ccctcagataggggtcccttgaccaccatcctccgtgaaatcaatatccc
gcacaagagtgctactctcctcgctccgggcccataacacttgggggtag
ctaaagtgaactgtatccgacatctggttcctacttcagggTcataaagc
ctaaatagcccacacgttccccttaaataagacatcacgatggatcacag
gtctatcaccctattaaccactcacgggagctctccatgcatttggtatt
ttcgtctggggggtatgcacgcgatagcattgcgagacgctggagccgga
gcaccctatgtcgcagtatctgtctttgattcctgcctcatccTattatt
tatcgcacctacgttcaatattacaggcgaacatacttactaaagtgtgt
taattaattaatgcttgtaggacataataataacaattgaatgtctgcac
agccActttccacacagacatcataacaaaaaatttccaccaaacccccc
XctcccccgcttctggccacagcacttaaacacatcTctgccaaaccccX
```

Fig. 17

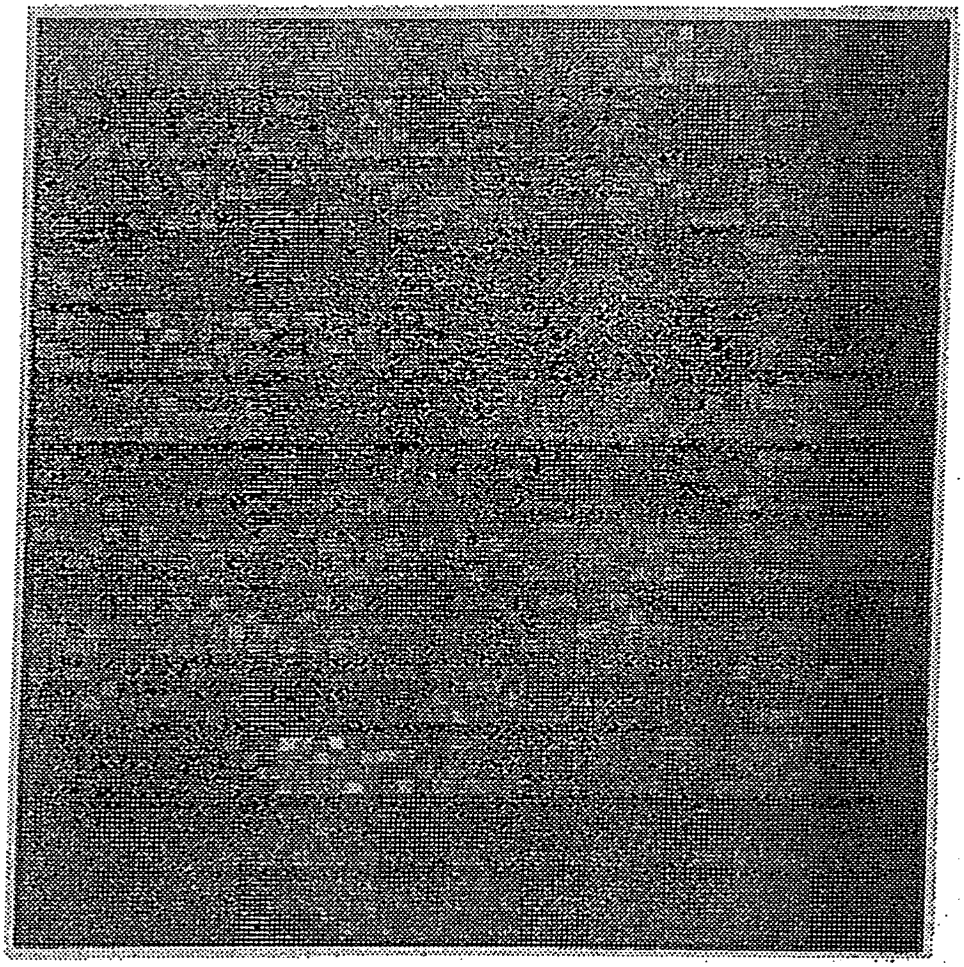
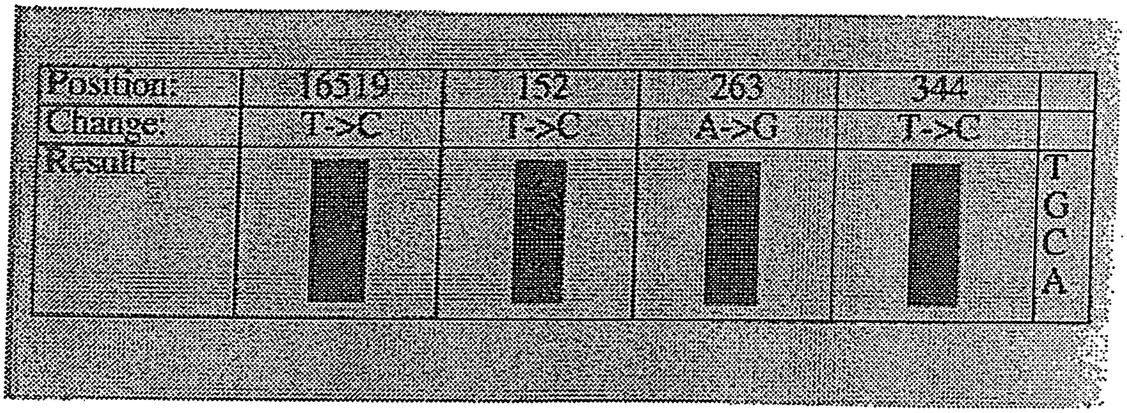


FIG. 18



Position:	16519	152	263	344	
Change:	T->C	T->C	A->G	T->C	
Result:	▮	▮	▮	▮	T G C A

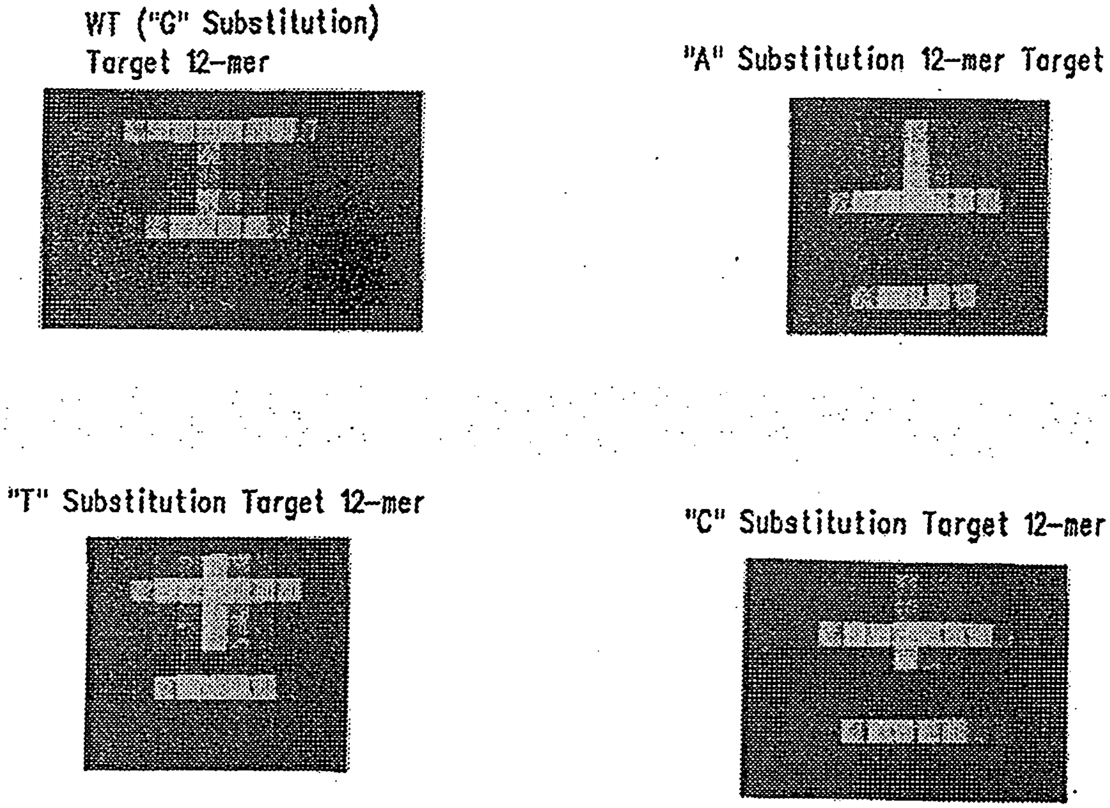
FIG. 19

```
                G A T G C T G A G G A G
0.                          C T C C T C C C C G G T
1.                        A C T C C T C C C C G G
2.                      G A C T C C T C C C C G
3.                    C G A C T C C T C C C C
4.                  A C G A C T C C T C C C
5.                T A C G A C T C C T C C
6.              C T A C G A C T C C T C
7.            T C T A C G A C T C C T
8.          T T C T A C G A C T C C
9.        A T T C T A C G A C T C
10.     T A T T C T A C G A C T
11.   C T A T T C T A C G A C
12. C C T A T T C T A C G A
```

Fig. 20

```
            G A T G C T G A G G A G
0.                        T C C T C C C C G G
1.                      C T C C T C C C C G
2.                    A C T C C T C C C C
3.                  G A C T C C T C C C
4.                C G A C T C C T C C
5.              A C G A C T C C T C
6.            T A C G A C T C C T
7.          C T A C G A C T C C
8.        T C T A C G A C T C
9.      T T C T A C G A C T
10.   A T T C T A C G A C
11. T A T T C T A C G A
```

Fig. 21



WT ("G" Substitution)
Target 12-mer

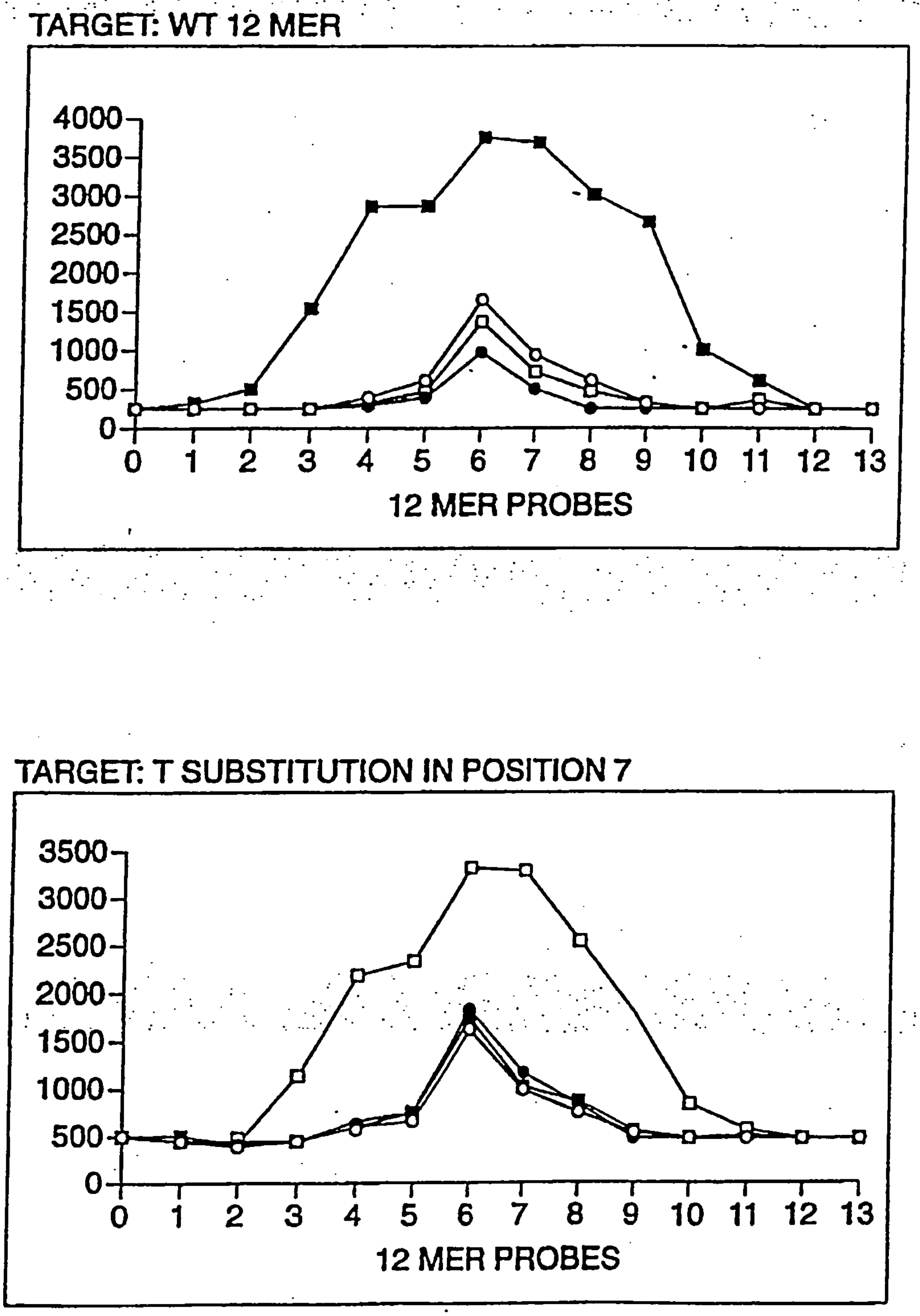
"A" Substitution 12-mer Target

"T" Substitution Target 12-mer

"C" Substitution Target 12-mer

FIG. 22

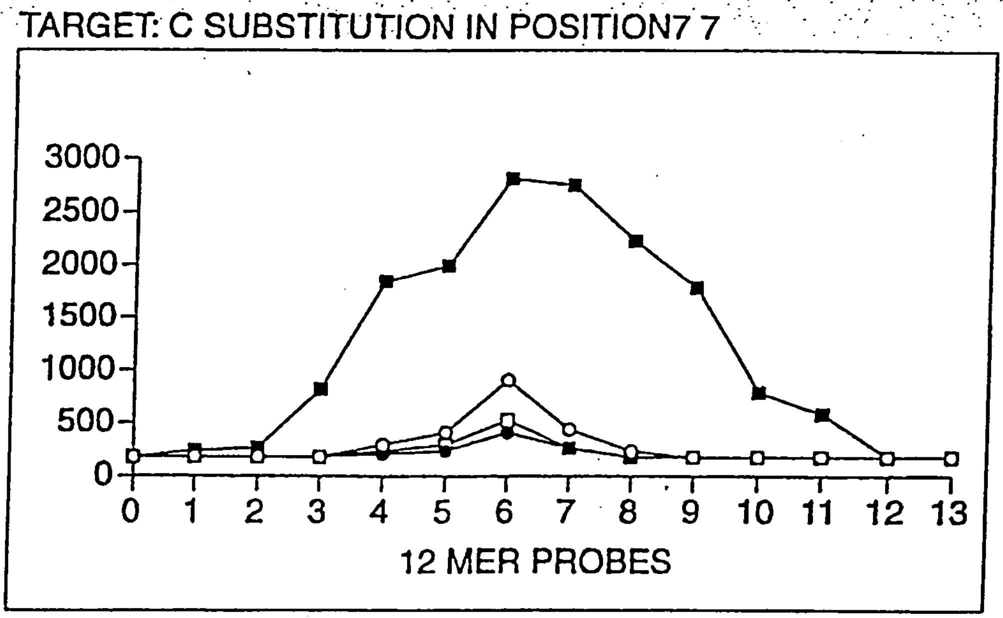
Fig. 23A

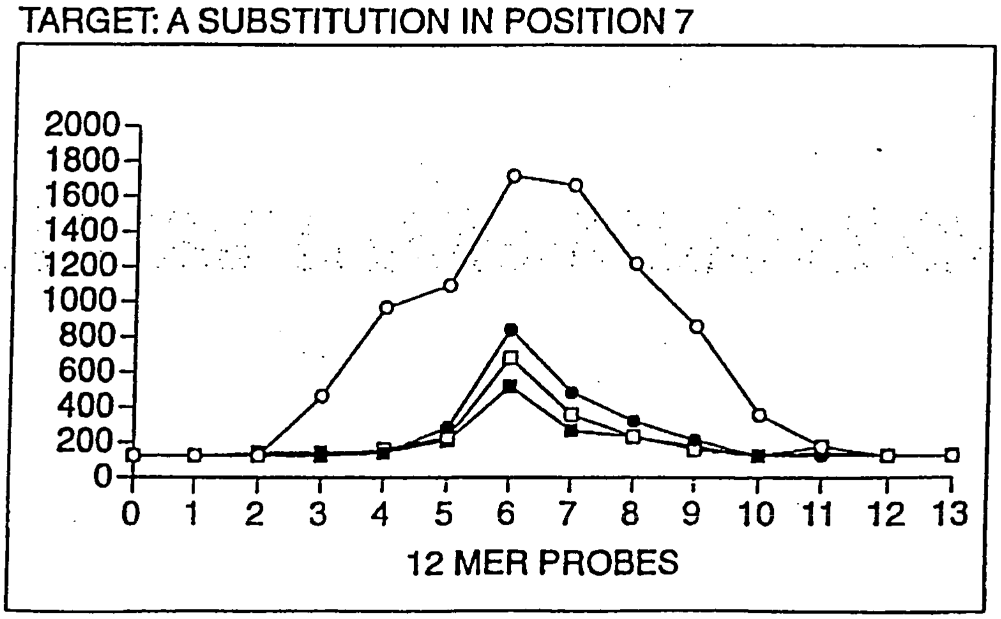


TARGET: WT 12 MER

TARGET: T SUBSTITUTION IN POSITION 7

Fig. 23B


TARGET: C SUBSTITUTION IN POSITION7 7
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
12 MER PROBES


TARGET: A SUBSTITUTION IN POSITION 7
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
12 MER PROBES

4:1 Mixture of WT and
"A" Substitution
12-mer Targets

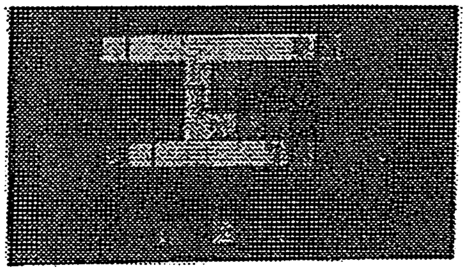
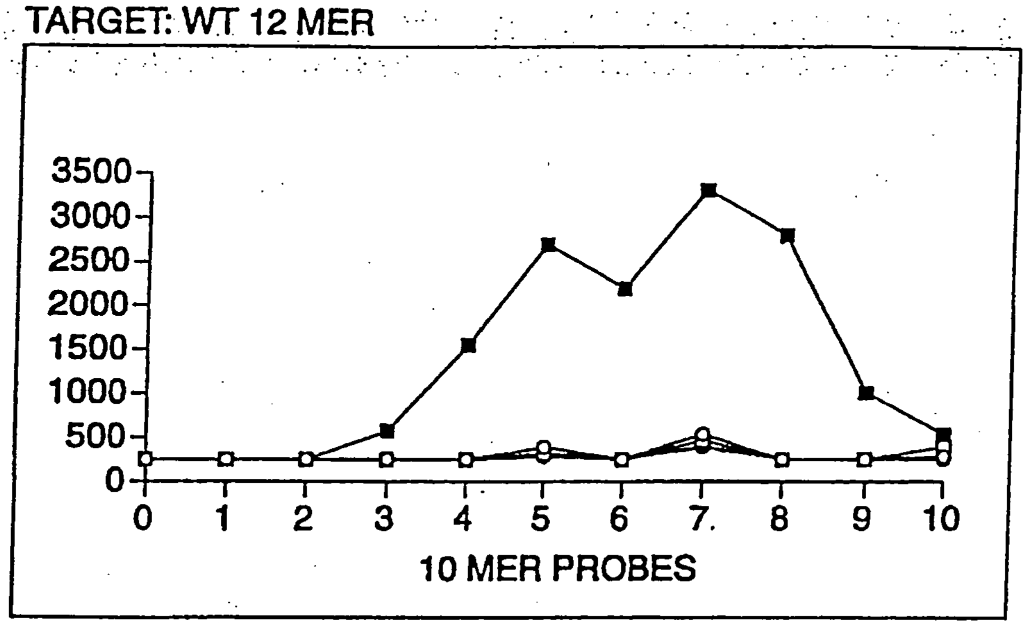


FIG. 24

Fig. 25A


TARGET: WT 12 MER
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
9
10
10 MER PROBES

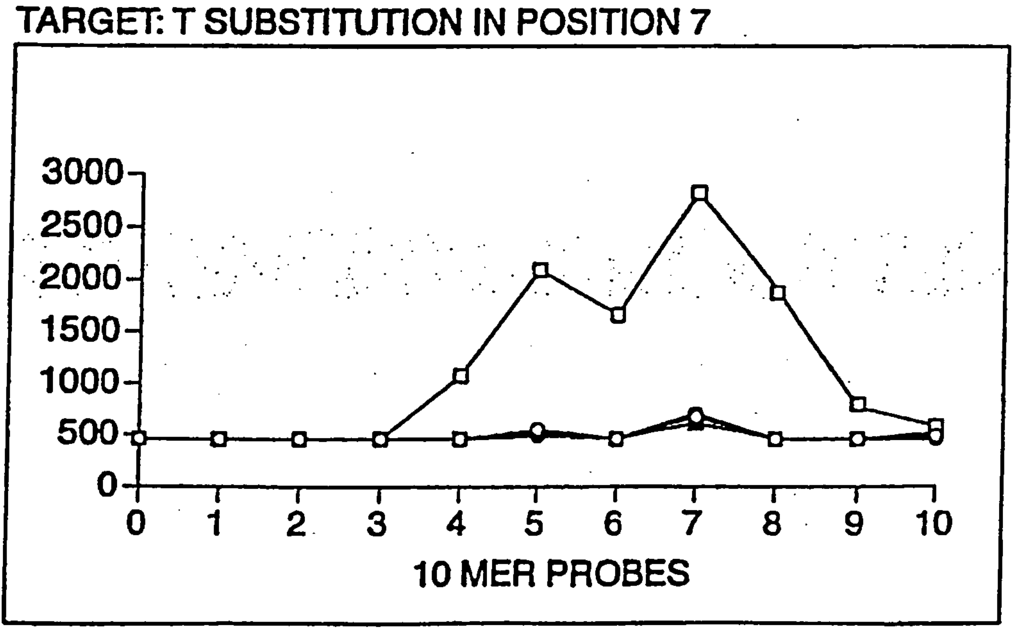
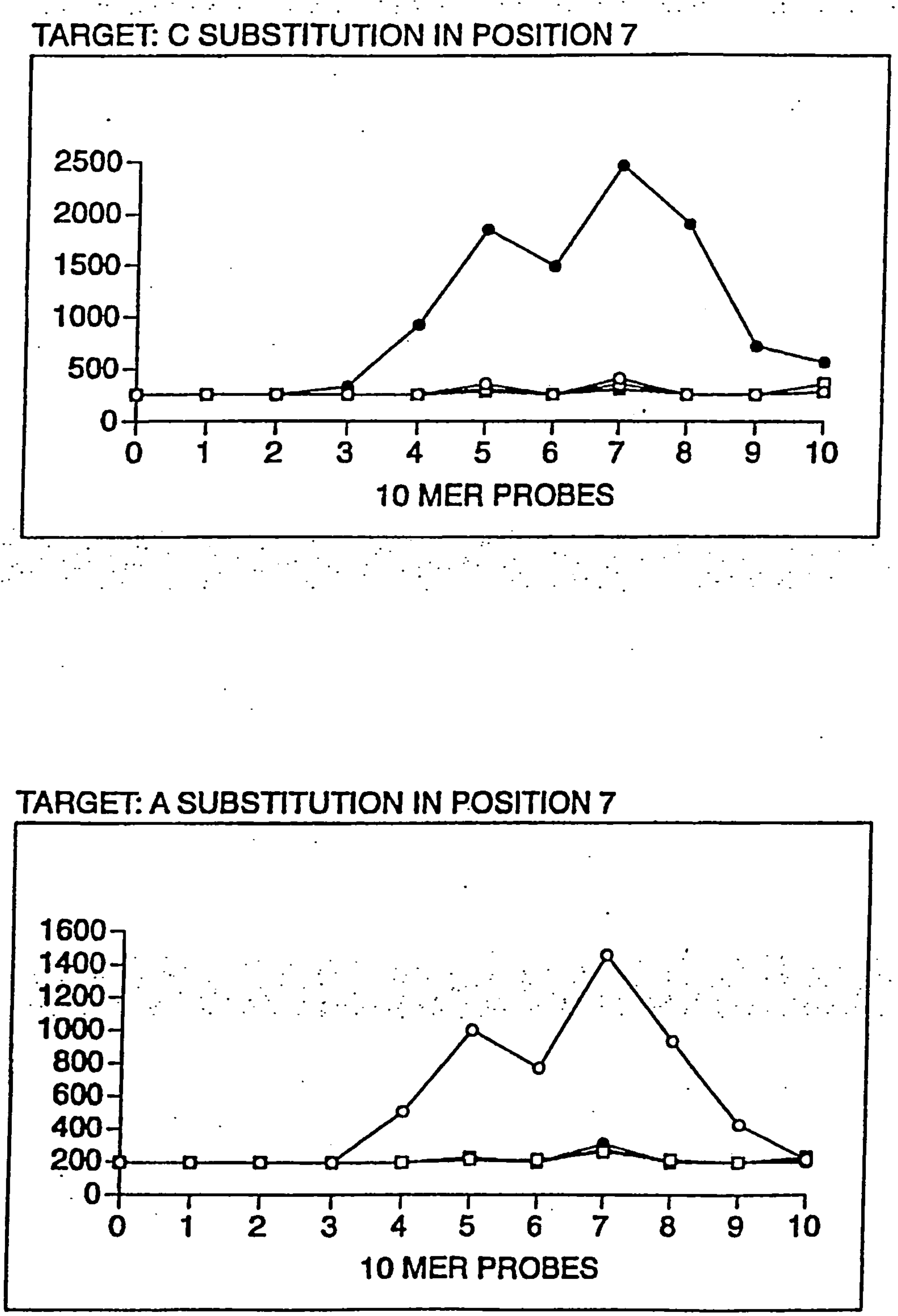

TARGET: T SUBSTITUTION IN POSITION 7
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
9
10
10 MER PROBES

Fig. 25B


TARGET: C SUBSTITUTION IN POSITION 7

TARGET: A SUBSTITUTION IN POSITION 7

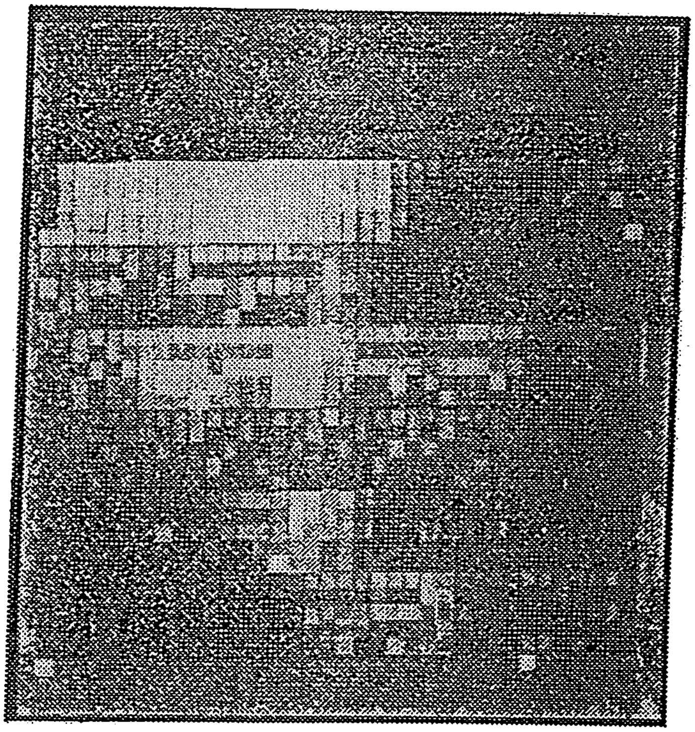


FIG. 26

WT	▨	T	A	C	T	C	C	C	C	T	G	C	C	C	T	C	A	A	C	A	A	G	A	T	G	T	T	T	T	G	C	C	A	A	C	▨
A	▨																																			▨
C	▨																																			▨
G	▨																																			▨
T	▨																																			▨
WT	T	G	G	C	C	A	A	G	A	C	C	T	G	C	C	C	T	G	T	G	C	A	G	C	T	G	T	G	G	G	T	T	G	A	T	T
A																									A						A	A				
C																																				
G																											G									
T																																				
WT	C	C	A	C	A	C	C	C	C	C	G	C	C	C	G	G	C	A	C	C	C	G	C	G	T	C	C	G	C	G	C	C	A	T	G	G
A																																				
C																																				
G																																				
T																																				
WT	C	C	A	T	C	T	A	C	A	A	G	C	A	G	T	C	A	▨	C	A	G	C	A	C	A	T	G	A	C	G	G	A	G	G	T	T
A																		▨																	A	A
C																		▨																		
G																		▨																		
T																		▨																		
WT	G	T	G	A	G	G	C	G	C	T	G	C	C	C	C	C	A	C	C	A	T	G	A	G	C	G	C	T	G	C	T	C	A	G	A	T
A																																				
C																												C			C					
G		G																																		
T																																				
WT	A	G	C	G	A	T	G	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
A								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
C				C		C		▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
G								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
T								▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

Fig. 27

Fig. 28

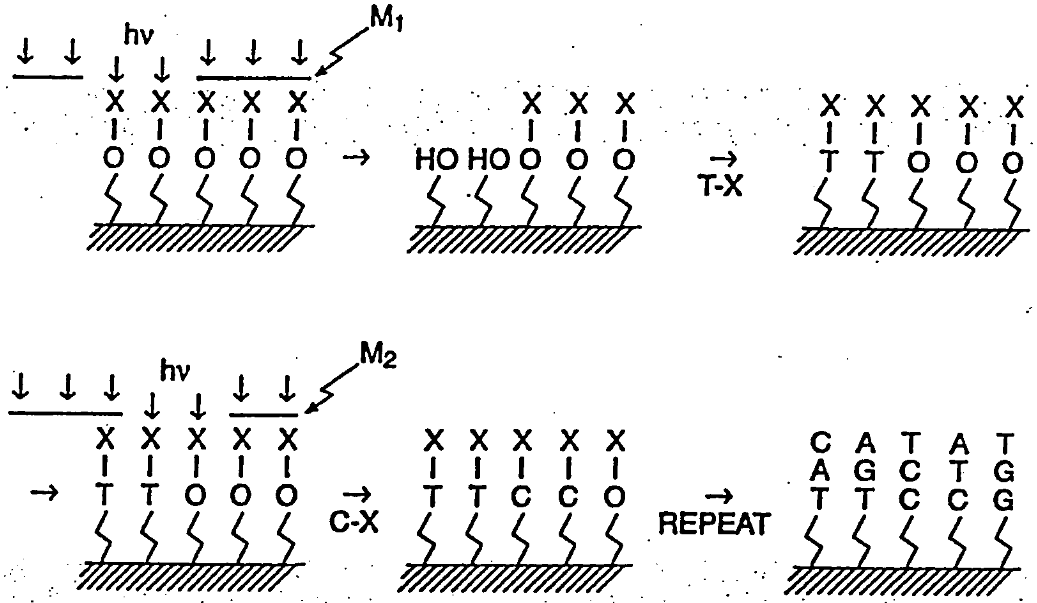
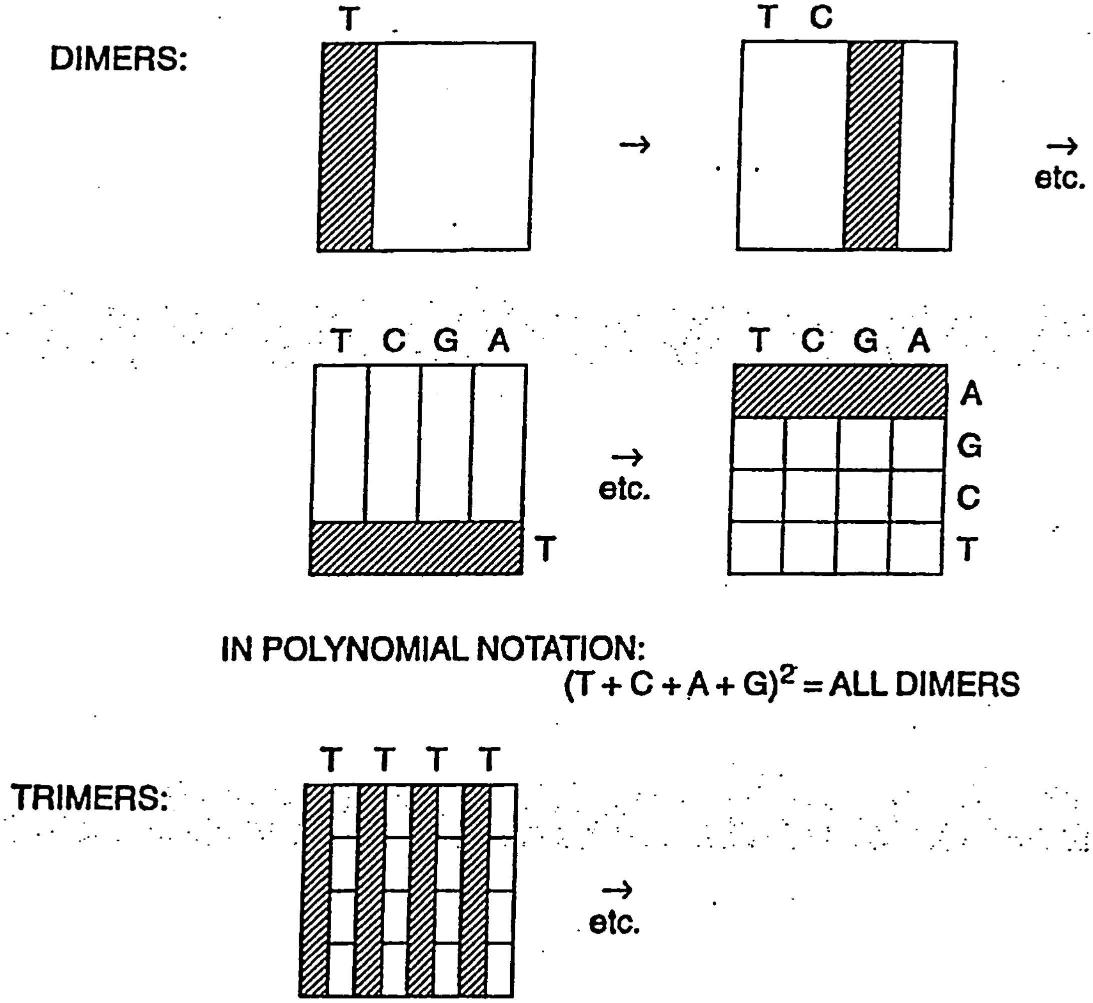

hν
M_1
X X X X X
O O O O O
HO HO O O O
T-X
T T O O O
hν
M_2
T T O O O
C-X
T T C C O
REPEAT
CAT AGT TCC ATC TGG

Fig. 29



IN POLYNOMIAL NOTATION:
$(T + C + A + G)^2$ = ALL DIMERS

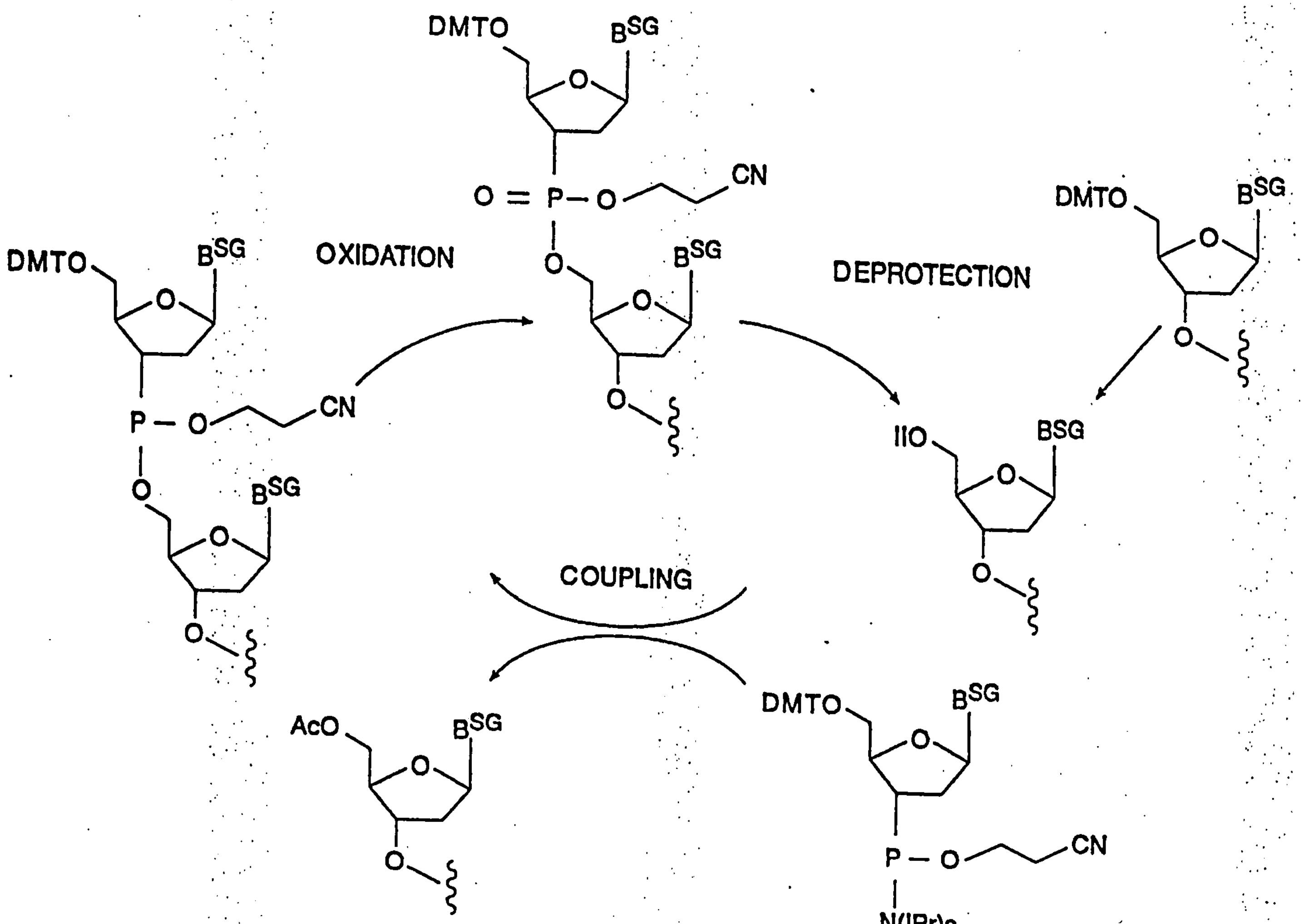

DMTO
B^{SG}
OXIDATION
O = P — O
CN
DEPROTECTION
P — O
IIO
BSG
COUPLING
AcO
N(IPr)2

Fig. 30

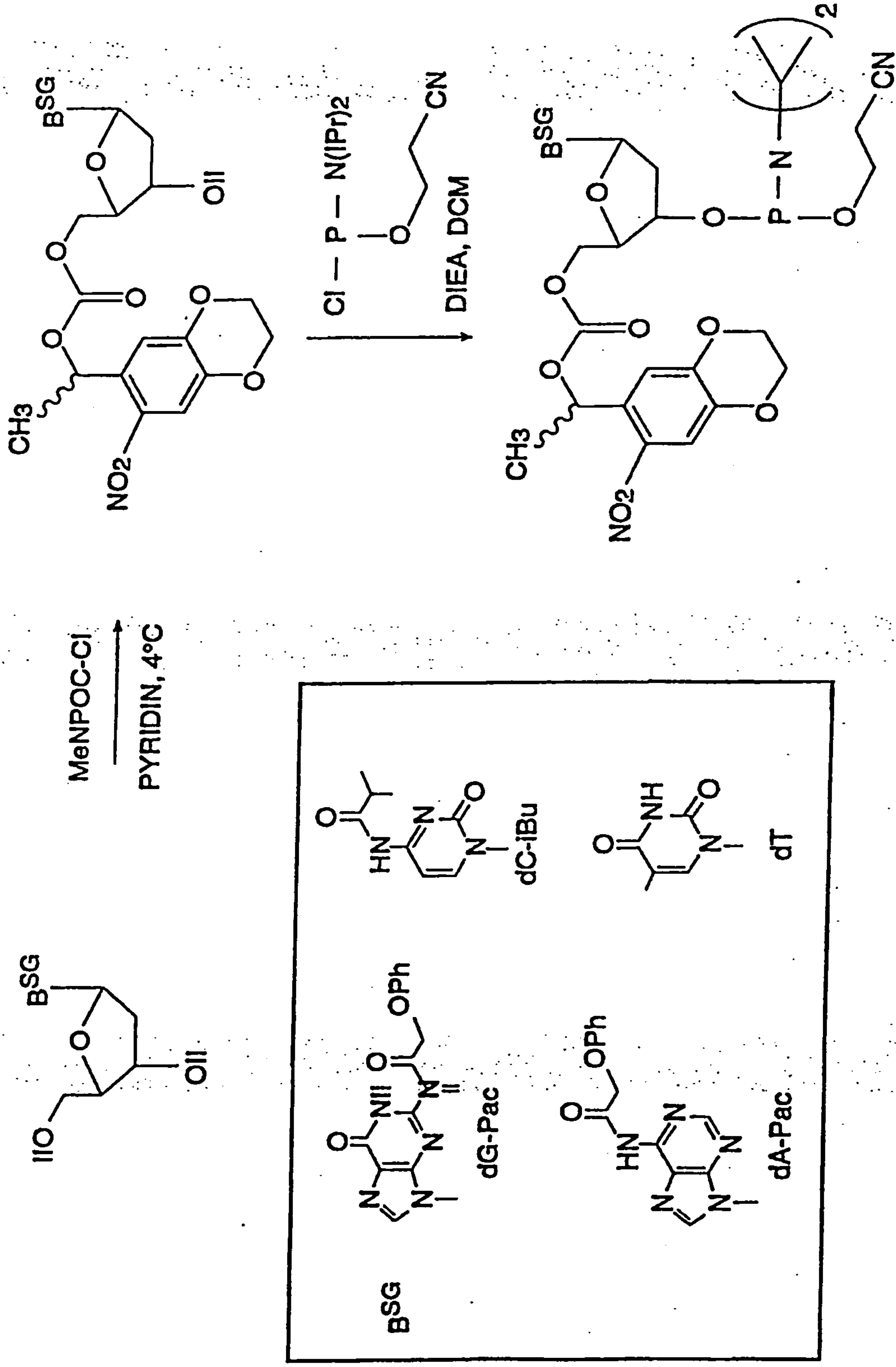

HO
OH
B^{SG}
MeNPOC-Cl
PYRIDIN, 4°C
CH3
NO2
B^{SG}
OH
Cl — P — N(iPr)2
CN
DIEA, DCM
CH3
NO2
B^{SG}
2
CN
B^{SG}
dG-Pac
OPh
dA-Pac
OPh
dC-iBu
dT

Fig. 31

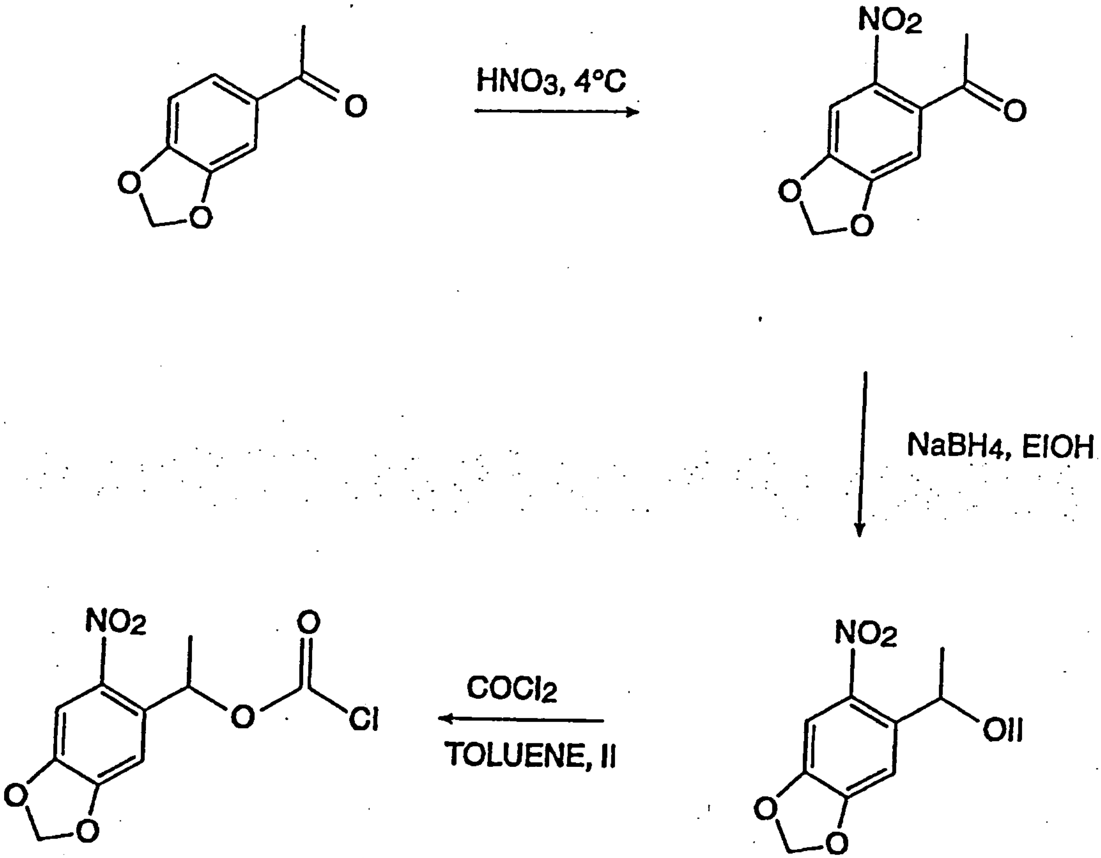


Fig. 32

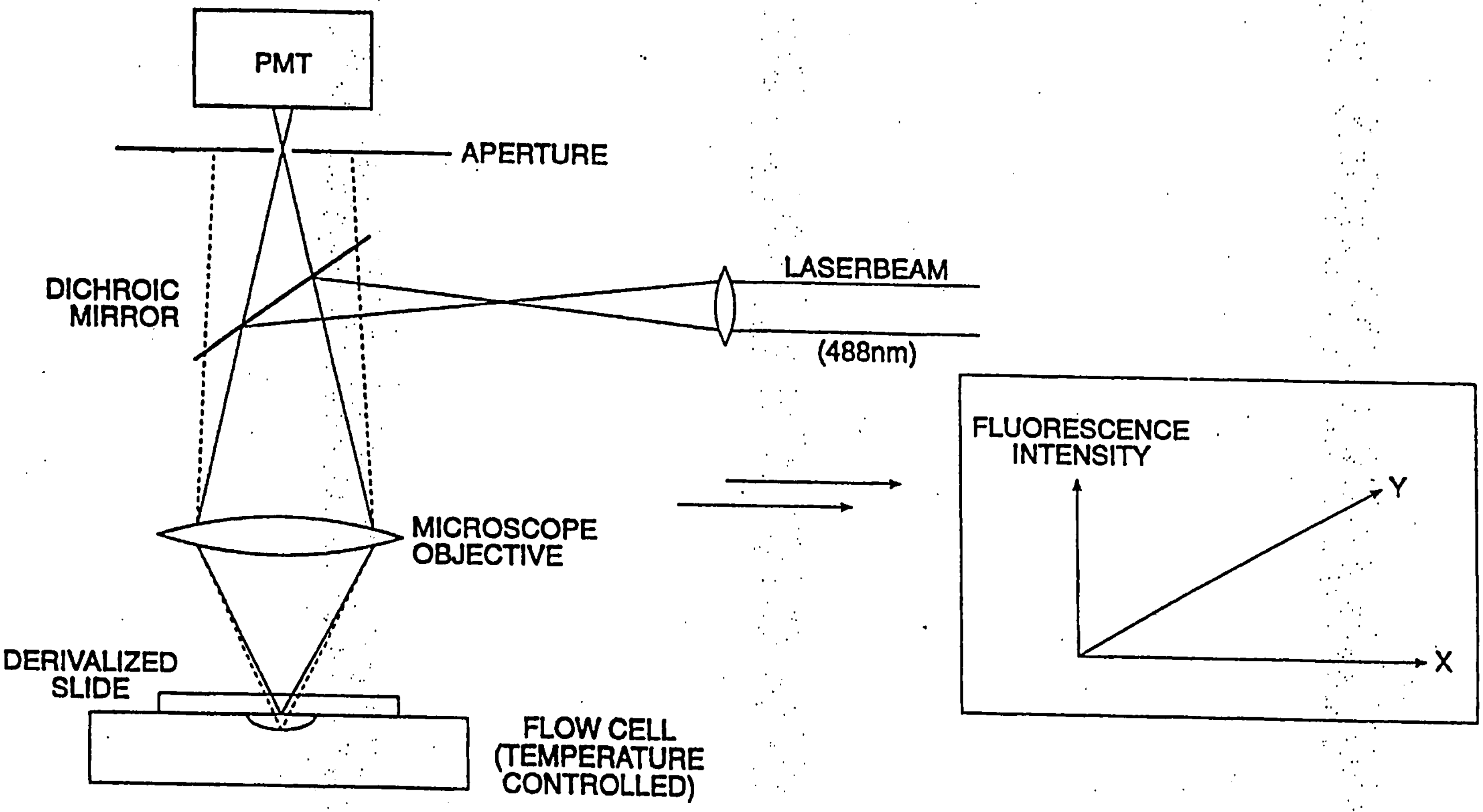

PMT
APERTURE
DICHROIC MIRROR
LASERBEAM
(488nm)
MICROSCOPE OBJECTIVE
DERIVALIZED SLIDE
FLOW CELL (TEMPERATURE CONTROLLED)
FLUORESCENCE INTENSITY
Y
X

Fig. 33

ARRAYS OF NUCLEIC ACID PROBES ON BIOLOGICAL CHIPS

CROSS-REFERENCE TO RELATED APPLICATION

This is a Continuation of application Ser. No. 08/143,312, filed Oct. 26, 1993, now abandoned, which is a continuation in part of U.S. patent application Ser. No. 082,937, filed 25 Jun. 1993, now abandoned, incorporated herein by reference.

Research leading to the invention was funded in part by NIH grant No. 1R01HG00813-01 and DOE grant No. DE-FG03-92-ER81275, and the government may have certain rights to the invention.

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention provides arrays of oligonucleotide probes immobilized in microfabricated patterns on silica chips for analyzing molecular interactions of biological interest. The invention therefore relates to diverse fields impacted by the nature of molecular interaction, including chemistry, biology, medicine, and medical diagnostics.

2. Description of Related Art

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid). In some assay formats, the oligonucleotide probe is tethered, i.e., by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. See, e.g., PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic but failed to provide an enabling method for using arrays of immobilized probes for this purpose. See U.S. Pat. Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126.

The development of VLSIPS™ technology has provided methods for making very large arrays of oligonucleotide probes in very small arrays. See U.S. Pat. No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated herein by reference. U.S. patent application Ser. No. 082,937, filed Jun. 25, 1993, describes methods for making arrays of oligonucleotide probes that can be used to provide the complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific nucleotide sequence.

Microfabricated arrays of large numbers of oligonucleotide probes, called "DNA chips" offer great promise for a wide variety of applications. New methods and reagents are required to realize this promise, and the present invention helps meet that need.

SUMMARY OF THE INVENTION

The present invention provides methods for making high-density arrays of oligonucleotide probes on silica chips and for using those probe arrays to detect specific nucleic acid sequences contained in a target nucleic acid in a sample. The invention also provides arrays of oligonucleotide probes on DNA chips, in which the probes have specific sequences and locations in the array to facilitate identification of a specific target nucleic acid. In another aspect, the invention provides methods for detecting whether one or more specific sequences of a target nucleic acid in a sample varies from a

previously characterized sequence or reference sequence. The methods of the invention can be used to detect variations between a target and reference sequence, including single or multiple base substitutions, and deletions and insertions of bases, as well as detecting the presence, location, and sequence of other more complex variations between a target and reference sequence in a nucleic acid.

The present invention provides arrays of oligonucleotide probes immobilized on a solid support. The arrays are preferably synthesized directly on the support using VLSIPS™ technology, but other synthesis methods and immobilization of pre-synthesized oligonucleotide probes can be used to make the oligonucleotide probe arrays, called "DNA chips", of the invention. In general, these arrays comprise a set of oligonucleotide probes such that, for each base in a specific reference sequence, the set includes a probe (called the "wild-type" or "WT" probe) that is exactly complementary to a section of the reference sequence including the base of interest and four additional probes (called "substitution probes"), which are identical to the WT probe except that the base of interest has been replaced by one of a predetermined set (typically 4) of nucleotides. In the preferred embodiment, one of the four substitution probes is identical to the wild type probe; the other three are complementary to targets that have a single-base substitution at this position.

In another aspect, the invention relates to the arrangement of individual probes in the array. In one embodiment, the probes are arranged on the chip so that probes for a given position in the sequence are adjacent, and probes for adjacent positions in the reference sequence are also adjacent to one another on the chip. One method arranges the probes for a single base in a short column (alternately row) and arranges the columns in the order of the base position to form horizontal (alternately vertical) stripes. The wild-type and each of the substitution probes have specified positions within the column so that all the probes corresponding to an A substitution, for example, are in a single row. The stripes may be separated on the chip by a blank row or column.

The DNA chips of the invention can be made in a wide number of variations. For some applications, leaving out the wild-type row, leaving out unimportant bases, pooling bases, including insertion and deletion probes, varying the length of the probes within a set to make the probes have the same or similar Tm relative to the target or to avoid secondary structure, varying the mutation position, using multiple probes for a single mutation, providing replicate probes or arrays, placing blank "streets" (no probe) between rows, columns, or individual probes, and using control probes may be appropriate.

The present invention also provides DNA chips for detecting mutations associated with cystic fibrosis, including mutations in exons 4, 7, 9, 10, 11, 20, and 21 of the CFTR gene. The invention also provides DNA chips for detecting mutations in the p53 gene, a gene in which mutations are known to be associated with a wide variety of cancers. Other DNA chips of the invention provide probe arrays for detecting specific sequences of mitochondrial DNA, useful for identification and forensic purposes. The invention also provides DNA chips for detecting specific sequences of nucleotides or mutations associated with the acquisition of a drug resistant phenotype in an infectious organism, such as rifampicin or other drug resistant TB strains and HIV, in which mutations in an RNA polymerase gene are known to give rise to drug resistance.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows how the tiling method of the invention defines a set of DNA probes relative to a target nucleic acid.

In the figure, the target is a DNA molecule, the probes are single-stranded nucleic acids 16 nucleotides in length, and only a portion of the probes defined by the method is shown.

FIG. 2 shows an illustrative tiled array of the invention with probes for the detection of point mutations. The base at the position of substitution in each of the wild-type probes is shown in the wild-type lane, and the shading shows the location of the substitution probe having the wild-type sequence. The SEQ ID. NOS. corresponding to the two peptide sequences shown in the top portion of FIG. 2 are 311 and 312, respectively. The SEQ ID. NOS. corresponding to the five peptide sequences listed at the bottom of FIG. 2 are 313, 314, 315, 313, and 316, respectively.

FIG. 3, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to a wild-type target; in panel C, the chip was hybridized to a mutant ΔF508 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets. The SEQ ID. NOS. corresponding to the four peptide sequences shown in FIG. 3 are 317–320, respectively.

FIG. 4, in sheets 1–3, corresponding to panels A, B, and C of FIG. 3, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest"). The SEQ ID. NOS. corresponding to the two peptide sequences shown in sheet 1 of FIG. 4 are 321 and 318, respectively; the SEQ ID. NOS. corresponding to the two peptide sequences shown in sheet 2 of FIG. 4 are 322 and 318, respectively; and the SEQ ID. NOS. corresponding to the two peptide sequences shown in sheet 3 of FIG. 4 are 323 and 318, respectively.

FIG. 5, in panels A, B, and C, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to the wt480 target; in panel C, the chip was hybridized to the mu480 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets. The SEQ ID. NOS. corresponding to the peptide sequences shown in FIG. 5 are 324–327, respectively.

FIG. 6, in sheets 1–3, corresponding to panels A, B, and C of FIG. 5, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest"). The SEQ ID. NOS. corresponding to the two peptide sequences shown in sheet 1 of FIG. 6 are 328 and 329, respectively; the SEQ ID. NOS. corresponding to the two peptide sequences shown in sheet 2 of FIG. 6 are 330 and 329, respectively; and the SEQ ID. NOS. corresponding to the two peptide sequences shown in sheet 3 of FIG. 6 are 331 and 329, respectively.

FIG. 7, in panels A and B, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to nucleic acid derived from the genomic DNA of an individual with wild-type ΔF508 sequences; in panel B, the target nucleic acid originated from a heterozygous (with respect to the ΔF508 mutation) individual.

FIG. 8, in sheets 1 and 2, corresponding to panels A and B of FIG. 7, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest"). The SEQ ID NOS. corresponding to the two peptide sequences shown in sheet 2 of FIG. 8 are 332 and 318, respectively.

FIG. 9 shows the human mitochondrial genome; "O_H" is the H strand origin of replication, and arrows indicate the cloned unshaded sequence.

FIG. 10 shows the image observed from application of a sample of mitochondrial DNA derived nucleic acid (from the mt4 sample) on a DNA chip.

FIG. 11 is similar to FIG. 10 but shows the image observed from the mt5 sample.

FIG. 12 shows the predicted difference image between the mt4 and mt5 samples on the DNA chip based on mismatches between the two samples and the reference sequence.

FIG. 13 shows the actual difference image observed for the mt4 and mt5 samples.

FIG. 14, in sheets 1 and 2, shows a plot of normalized intensities across rows 10 and 11 of the array and a tabulation of the mutations detected.

FIG. 15 shows the discrimination between wild-type and mutant hybrids obtained with the chip. A median of the six normalized hybridization scores for each probe was taken; the graph plots the ratio of the median score to the normalized hybridization score versus mean counts. A ratio of 1.6 and mean counts above 50 yield no false positives.

FIG. 16 illustrates how the identity of the base mismatch may influence the ability to discriminate mutant and wild-type sequences more than the position of the mismatch within an oligonucleotide probe. The mismatch position is expressed as % of probe length from the 3'-end. The base change is indicated on the graph.

FIG. 17 provides a 5' to 3' sequence listing of one target corresponding to the probes on the chip. X is a control probe. Positions that differ in the target (i.e., are mismatched with the probe at the designated site) are in bold. The SEQ ID. NO. corresponding to the peptide sequence shown in FIG. 17 is 333.

FIG. 18 shows the fluorescence image produced by scanning the chip described in FIG. 17 when hybridized to a sample.

FIG. 19 illustrates the detection of 4 transitions in the target sequence relative to the wild-type probes on the chip in FIG. 18.

FIG. 20 shows the alignment of some of the probes on a p^{53} DNA chip with a 12-mer model target nucleic acid. The SEQ ID. NOS. corresponding to the fourteen peptide sequences shown in FIG. 20 are 334–347, respectively.

FIG. 21 shows a set of 10-mer probes for a p53 exon 6 DNA chip. The SEQ ID. NOS. corresponding to the thirteen peptide sequences shown in FIG. 21 are 334 and 348–359, respectively.

FIG. 22 shows that very distinct patterns are observed after hybridization of p53 DNA chips with targets having different 1 base substitutions. In the first image in FIG. 22, the 12-mer probes that form perfect matches with the wild-type target are in the first row (top). The 12-mer probes with single base mismatches are located in the second, third, and fourth rows and have much lower signals.

FIG. 23, in graphs 2, 3, and 4, graphically depicts the data in FIG. 22. On each graph, the X ordinate is the position of the probe in its row on the chip, and the Y ordinate is the signal at that probe site after hybridization.

FIG. 24 shows the results of hybridizing mixed target populations of WT and mutant p53 genes to the p53 DNA chip.

FIG. 25, in graphs 1–4, shows (see FIG. 23 as well) the hybridization efficiency of a 10-mer probe array as compared to a 12-mer probe array

FIG. 26 shows an image of a p53 DNA chip hybridized to a target DNA.

FIG. 27 illustrates how the actual sequence was read from the chip shown in FIG. 26. Gaps in the sequence of letters in the WT rows correspond to control probes or sites. Positions at which bases are miscalled are represented by letters in italic type in cells corresponding to probes in which the WT bases have been substituted by other bases. The SEQ ID. NO. corresponding to the peptide sequence shown in FIG. 27 is 360.

FIG. 28 illustrates the VLSIPS™ technology as applied to the light directed synthesis of oligonucleotides. Light (hν) is shone through a mask (M_1) to activate functional groups (—OH) on a surface by removal of a protecting group (X). Nucleoside building blocks protected with photoremovable protecting groups (T-X, C-X) are coupled to the activated areas. By repeating the irradiation and coupling steps, very complex arrays of oligonucleotides can be prepared.

FIG. 29 illustrates how the VLSIPS™ process can be used to prepare "nucleoside combinatorials" or oligonucleotides synthesized by coupling all four nucleosides to form dimers, trimers, etc.

FIG. 30 shows the deprotection, coupling, and oxidation steps of a solid phase DNA synthesis method.

FIG. 31 shows an illustrative synthesis route for the nucleoside building blocks used in the VLSIPS™ method.

FIG. 32 shows a preferred photoremovable protecting group, MeNPOC, and how to prepare the group in active form.

FIG. 33 illustrates an illustrative detection system for scanning a DNA chip.

DETAILED DESCRIPTION OF THE INVENTION

Using the VLSIPS™ method, one can synthesize arrays of many thousands of oligonucleotide probes on a substrate, such as a glass slide or chip. The method can be used, for instance, to synthesize "combinatorial" arrays consisting of, for example, all possible octanucleotides. Such arrays can be used for primary sequencing-by-hybridization on genomic DNA fragments or other nucleic acids or to detect mutations in a target nucleic acid for which the normal or "wild-type" nucleotide sequence is already known. Using the preferred method of the invention, one employs a strategy called "tiling" to synthesize specific sets of probes or at spatially-defined locations on a substrate, creating the novel probe arrays and "DNA chips" of the invention.

To illustrate the tiling method of the invention, consider the problem of detecting mutations at one or more position in the nucleotide sequence of a target nucleic acid with oligonucleotide probes of defined length. The length (L) of the probe is typically expressed as the number of nucleotides or bases in a single-stranded nucleic acid probe. For purposes of the present invention, lengths ranging from 12 to 18 bases are preferred, although shorter and longer lengths can also be employed. To employ the tiling method, one synthesizes a set of probes defined by the particular nucleotide sequence of interest in the target nucleic acid. For each base in the target DNA segment, one synthesizes a probe complementary to the subsequence of the target nucleic acid beginning at that base and ending L-1 bases to the 3'-side (see FIG. 1).

In a preferred embodiment of the invention, the probes are arranged (either by immobilization, typically by covalent attachment, of a pre-synthesized probe or by synthesis of the probe on the substrate) on the substrate or chips in lanes stretching across the chip and separated, and these lanes are in turned arranged in blocks of preferably 5 lanes, although blocks of other sizes will have useful application, as will be apparent from the following illustration. The first of these five lanes, called the "wild-type lane", contains probes arranged in order of sequence, and all of the probes are complementary to a specified wild-type nucleic acid sequence. The other four lanes contain probe sets for detecting all possible single-base mutations in the defined sequence; in turn, these probe sets are defined by a position of potential non-complementarity in the probe relative to the target (i.e., a single base mismatch) and the identity of the nucleotide in the probe at that position (i.e., whether the nucleotide is an A, C, G, or T nucleotide). The position of mismatch, also called the position of substitution, is preferably selected to be near the center of the probes, i.e., position 7 of a probe of L=15.

For each probe in the wild-type lane, one synthesizes four probes (one for each of the lanes other than the wild-type lane), Three of these four probes is identical to the corresponding wild-type probe but for the base at the position of substitution, and the remaining probe is identical to the wild-type probe. This set of four substitution probes is preferably placed in a column directly below (or above) the corresponding wild-type probe, thus creating an A-lane, a C-lane, a G-lane, and a T-lane. FIG. 2 shows an illustrative tiled array of the invention with probes for the detection of point mutations. The base at the position of substitution in each of the wild-type probes is shown in the wild-type lane, and the shading shows the location of the substitution probe having the wild-type sequence. Below are the probes that would be placed in the column marked by the arrow if the probe length were 15 and the position of substitution were 7.

3'-CCGACTGCAGTCGTT (SEQ. ID. NO:1)
3'-CCGACTACAGTCGTT (SEQ. ID. NO:2)
3'-CCGACTCCAGTCGTT (SEQ. ID. NO:3)
3'-CCGACTGCAGTCGTT (SEQ. ID. NO:1)
3'-CCGACTTCAGTCGTT (SEQ. ID. NO:4)

Thus, the substitution lanes occupy four of the five lanes separating successive wild-type lanes on the chip; the blocks of five lanes can be separated by a sixth lane for measurement of background signals.

The DNA chips of the invention have a wide variety of applications. In one embodiment, the DNA chip is used to select an optimal probe from an array of probes. In this embodiment, an array of probes of variable length and sequences is synthesized and then hybridized to a target nucleic acid of known sequence. The pattern of hybridization reveals the optimal length and sequence composition of probes to detect a particular mutation or other specific sequence of nucleotides. In some circumstances, i.e., target nucleic acids with repeated sequences or with high G/C content, very long probes may be required for optimal detection. In one embodiment for detecting specific sequences in a target nucleic acid with a DNA chip, repeat sequences are detected as follows. The chip comprises probes of length sufficient to extend into the repeat region varying distances from each end. The sample, prior to hybridization, is treated with a labeled oligonucleotide that is complementary to a repeat region but shorter than the full length of the repeat. The target nucleic is labeled with a second, distinct label. After hybridization, the chip is scanned for probes that have bound both the labeled target and the labeled oligonucleotide probe; the presence of such bound probes shows that at least two repeat sequences are present.

A variety of methods can be used to enhance detection of labeled targets bound to a probe on the array. In one embodiment, the protein MutS (from *E. coli*) or equivalent proteins such as yeast MSH1, MSH2, and MSH3; mouse Rep-3, and Streptococcus Hex-A, is used in conjunction with target hybridization to detect probe-target complex that contain mismatched base pairs. The protein, labeled directly or indirectly, can be added to the chip during or after hybridization of target nucleic acid, and differentially binds to homo- and heteroduplex nucleic acid. A wide variety of dyes and other labels can be used for similar purposes. For instance, the dye YOYO-1 is known to bind preferentially to nucleic acids containing sequences comprising runs of 3 or more G residues.

The DNA chips produced by the methods of the invention can be used to study and detect mutations in exons of human genes of clinical interest, including point mutations and deletions. In the following sections, the method of the invention is illustrated by the detection of mutations in a variety of clinically and medically significant human nucleic acid sequences. Thus, the invention is illustrated first with respect to the preparation of DNA chips for the detection of mutations associated with cystic fibrosis, then with DNA chips for the detection of human mitochondrial DNA sequences, then with DNA chips for the detection of mutations in the human p53 gene associated with cancer, and finally with respect to the detection of mutations in the HIV RT gene associated with drug resistance.

Detection of Cystic Fibrosis Mutations with DNA Chips

A number of years ago, cystic fibrosis, the most common severe autosomal recessive disorder in humans, was shown to be associated with mutations in a gene thereafter named the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. The sequences of the exons and parts of the introns in the gene are known, as are the changes corresponding to several hundred known mutations. Several tests have been developed for detecting the most frequent of these mutations. The present invention provides CFTR gene oligonucleotide arrays (DNA chips) that can be used to identify mutations in the CFTR gene rapidly and efficiently.

The methods used to make the high-density DNA chips of the invention allow probes for long stretches of DNA coding regions to be directly "written" onto the chips in the form of sets of overlapping oligonucleotides. These methods have been used to develop a number of useful CFTR gene chips, one illustrative chip bears an array of 1296 probes covering the full length of exon 10 of the CFTR gene arranged in a 36×36 array of 356 λm elements. The probes in the array can have any length, preferably in the range of from 10 to 18 residues and can be used to detect and sequence any single-base substitution and any deletion within the 192-base exon, including the three-base deletion known as ΔF508. As described in detail below, hybridization of sub-nanomolar concentrations of wild-type and ΔF508 oligonucleotide target nucleic acids labeled with fluorescein to these arrays produces highly specific signals (detected with confocal scanning fluorescence microscopy) that permit discrimination between mutant and wild-type target sequences in both homozygous and heterozygous cases. The method and chips of the invention can also be used to detect other known mutations in the CFTR gene, as described in detail below.

The most common cystic fibrosis mutation is known as ΔF508, because the mutation is a three-base deletion that results in the removal of amino acid #508 from the CFTR protein. The present invention provides DNA chips for detecting ΔF508, one such chip results from applying the tiling method to exon 10 of the CFTR gene, the exon to which ΔF508 has been mapped. The tiling method involved the synthesis of a set of probes of a selected length in the range of from 10 to 18 bases and complementary to subsequences of the known wild-type CFTR sequence starting at a position a few bases into the intron on the 5'-side of exon 10 and ending a few bases into the intron on the 3'-side. There was a probe for each possible subsequence of the given segment of the gene, and the probes were organized into a "lane" in such a way that traversing the lane from the upper left-hand corner of the chip to the lower righthand corner corresponded to traversing the gene segment base-by-base from the 5'-end. The lane containing that set of probes is, as noted above, called the "wild-type lane."

Relative to the wild-type lane, a "substitution" lane, called the "A-lane", was synthesized on the chip. The A-lane probes were identical in sequence to an adjacent (immediately below the corresponding) wild-type probe but contained, regardless of the sequence of the wild-type probe, a dA residue at position 7 (counting from the 3'-end). In similar fashion, substitution lanes with replacement bases dC, dG, and dT were placed onto the chip in a "C-lane," a "G-lane," and a "T-lane," respectively. A sixth lane on the chip consisted of probes identical to those in the wild-type lane but for the deletion of the base in position 7 and restoration of the original probe length by addition to the 5'-end the base complementary to the gene at that position.

The four substitution lanes enable one to deduce the sequence of a target exon 10 nucleic acid from the relative intensities with which the target hybridizes to the probes in the various lanes. The probe organization on the chip can be conveniently columnar, and the set of probes consisting of a wild-type probe and four corresponding substitution probes is referred to as a "column set." One and only one of the four substitution probes in a column set has exactly the same sequence as the wild-type probe in the set. Those of skill in the art will appreciate that, in other embodiments of the invention, one could delete one or more lanes or columns and still benefit from the invention. Various versions of such exon 10 DNA chips were made as described above with probes 15 bases long, as well as chips with probes 10, 14, and 18 bases long. For the results described below, the probes were 15 bases long, and the position of substitution was 7 from the 3'-end.

To demonstrate the ability of the chip to distinguish the ΔF508 mutation from the wild-type, two synthetic target nucleic acids were made. The first, a 39-mer complementary to a subsequence of exon 10 of the CFTR gene having the three bases involved in the ΔF508 mutation near its center, is called the "wild-type" or wt508 target, corresponds to positions 111–149 of the exon, and has the sequence shown below:

5'-CATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGA (SEQ. ID NO: 5).

The second, a 36-mer probe derived from the wild-type target by removing those same three bases, is called the "mutant" target or mu508 target and has the sequence shown below, first with dashes to indicate the deleted bases, and then without dashes but with one base underlined (to indicate the base detected by the T-lane probe, as discussed below):

5'-CATTAAAGAAAATATCAT---TGGTGTTTCCTATGATGA; (SEQ. ID NO:6)

5'-CATTAAAGAAAATATCATTGGTGTTTCCTATGATGA. (SEQ. ID NO:7)

Both targets were labeled with fluorescein at the 5'-end.

In three separate experiments, the wild-type target, the mutant target, and an equimolar mixture of both targets was exposed (0.1 nM wt508, 0.1 nM mu508, and 0.1 nM wt508 plus 0.1 nM mu508, respectively, in a solution compatible with nucleic acid hybridization) to a CF chip. The hybridization mixture was incubated overnight at room temperature, and then the chip was scanned on a reader (a confocal fluorescence microscope in photon-counting mode; images of the chip were constructed from the photon counts) at several successively higher temperatures while still in contact with the target solution. After each temperature change, the chip was allowed to equilibrate for approximately one-half hour before being scanned. After each set of scans, the chip was exposed to denaturing solvent and conditions to wash, i.e., remove target that had bound, the chip so that the next experiment could be done with a clean chip.

The results of the experiments are shown in FIGS. 3, 4, 5, and 6. FIG. 3, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to a wild-type target; in panel C, the chip was hybridized to a mutant delta 508 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets. FIG. 4, in sheets 1–3, corresponding to panels A, B, and C of FIG. 3, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

These figures show that, for the wild-type target and the equimolar mixture of targets, the substitution probe with a nucleotide sequence identical to the corresponding wild-type probe bound the most target, allowing for an unambiguous assignment of target sequence as shown by letters near the points on the curve. The target wt508 thus hybridized to the probes in the wild-type lane of the chip, although the strength of the hybridization varied from probe-to-probe, probably due to differences in melting temperature. The sequence of most of the target can thus be read directly from the chip, by inference from the pattern of hybridization in the lanes of substitution probes (if the target hybridizes most intensely to the probe in the A-lane, then one infers that the target has a T in the position of substitution, and so on).

For the mutant target, the sequence could similarly be called on the 3'-side of the deletion. However, the intensity of binding declined precipitously as the point of substitution approached the site of the deletion from the 3'-end of the target, so that the binding intensity on the wild-type probe whose point of substitution corresponds to the T at the 3'-end of the deletion was very close to background. Following that pattern, the wild-type probe whose point of substitution corresponds to the middle base (also a T) of the deletion bound still less target. However, the probe in the T-lane of that column set bound the target very well.

Examination of the sequences of the two targets reveals that the deletion places an A at that position when the sequences are aligned at their 3'-ends and that the T-lane probe is complementary to the mutant target with but two mismatches near an end (shown below in lower-case letters, with the position of substitution underlined):

Target: 5'-CATTAAAGAAAATATCATTGGTGTTTCCTATGATGA

Probe: 3'-TagTAGTAACCACAA (SEQ. ID NO:8)

Thus the T-lane probe in that column set calls the correct base from the mutant sequence. Note that, in the graph for the equimolar mixture of the two targets, that T-lane probe binds almost as much target as does the A-lane probe in the same column set, whereas in the other column sets, the probes that do not have wild-type sequence do not bind target at all as well. Thus, that one column set, and in particular the T-lane probe within that set, detects the ΔF508 mutation under conditions that simulate the homozygous case and also conditions that simulate the heterozygous case.

The present invention thus provides individual probes, sets of probes, and arrays of probe sets on chips, in specific patterns, as the probes provide important benefits for detecting the presence of specific exon 10 sequences. The sequences of several important probes of the invention are shown below. In each case, the letter "X" stands for the point of substitution in a given column set, so each of the sequences actually represents four probes, with A, C, G, and T, respectively, taking the place of the "X." Sets of shorter probes derived from the sets shown below by removing up to five bases from the 5'-end of each probe and sets of longer probes made from this set by adding up to three bases from the exon 10 sequence to the 5'-end of each probe, are also useful and provided by the invention.

3'-TTTATAXTAGAAACC (SEQ. ID NO:9)
3'-TTATAGXAGAAACCA (SEQ. ID NO:10)
3'-TATAGTXGAAACCAC (SEQ. ID NO:11)
3'-ATAGTAXAAACCACA (SEQ. ID NO:12)
3'-TAGTAGXAACCACAA (SEQ. ID NO:13)
3'-AGTAGAXACCACAAA (SEQ. ID NO:14)
3'-GTAGAAXCCACAAAG (SEQ. ID NO:15)
3'-TAGAAAXCACAAAGG (SEQ. ID NO:16)
3'-AGAAACXACAAAGGA (SEQ. ID NO:17)

Although in this example the sequence could not be reliably deduced near the ends of the target, where there is not enough overlap between target and probe to allow effective hybridization, and around the center of the target, where hybridization was weak for some other reason, perhaps high AT-content, the results show the method and the probes of the invention can be used to detect the mutation of interest. The mutant target gave a pattern of hybridization that was very similar to that of the wt508 target at the ends, where the two share a common sequence, and very different in the middle, where the deletion is located. As one scans the image from right to left, the intensity of hybridization of the target to the probes in the wild-type lane drops off much more rapidly near the center of the image for mu508 than for wt508; in addition, there is one probe in the T-lane that hybridizes intensely with mu508 and hardly at all with wt508. The results from the equimolar mixture of the two targets, which represents the case one would encounter in testing a heterozygous individual for the mutation, are a blend of the results for the separate targets, showing the power of the invention to distinguish a wild-type target sequence from one containing the ΔF508 mutation and to detect a mixture of the two sequences.

The results above clearly demonstrate how the DNA chips of the invention can be used to detect a deletion mutation, ΔF508; another model system was used to show that the chips can also be used to detect a point mutation as well. One of the more frequent mutations in the CFTR gene is G480C, which involves the replacement of the G in position 46 of exon 10 by a T, resulting in the substitution of a cysteine for the glycine normally in position #480 of the CFTR protein. The model target sequences included the 21-mer probe wt480 to represent the wild-type sequence at positions 37–55 of exon 10: 5'-CCTTCAGAGGGTAAAATTAAG (SEQ. ID NO:18) and the 21-mer probe mu480 to represent the mutant sequence: 5'-CCTTCAGAGTGTAAAATTAAG (SEQ. ID NO:19).

In separate experiments, a DNA chip was hybridized to each of the targets wt480 and mu480, respectively, and then scanned with a confocal microscope. FIG. 5, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to the wt480 target; in panel C, the chip was hybridized to the mu480 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets. FIG. 6, in sheets 1–3, corresponding to panels A, B, and C of FIG. 5, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

These figures show that the chip could be used to sequence a 16-base stretch from the center of the target wt480 and that discrimination against mismatches is quite good throughout the sequenced region. When the DNA chip was exposed to the target mu480, only one probe in the portion of the chip shown bound the target well: the probe in the set of probes devoted to identifying the base at position 46 in exon 10 and that has an A in the position of substitution and so is fully complementary to the central portion of the mutant target. All other probes in that region of the chip have at least one mismatch with the mutant target and therefore bind much less of it. In spite of that fact, the sequence of mu480 for several positions to both sides of the mutation can be read from the chip, albeit with much-reduced intensities from those observed with the wild-type target.

The results also show that, when the two targets were mixed together and exposed to the chip, the hybridization pattern observed was a combination of the other two patterns. The wild-type sequence could easily be read from the chip, but the probe that bound the mu480 target so well when only the mu480 target was present also bound it well when both the mutant and wild-type targets were present in a mixture, making the hybridization pattern easily distinguishable from that of the wild-type target alone. These results again show the power of the DNA chips of the invention to detect point mutations in both homo- and heterozygous individuals.

To demonstrate clinical application of the DNA chips of the invention, the chips were used to study and detect mutations in nucleic acids from genomic samples. Genomic samples from a individual carrying only the wild-type gene and an individual heterozygous for ΔF508 were amplified by PCR using exon 10 primers containing the promoter for T7 RNA polymerase. Illustrative primers of the invention are shown below.

Exon	Name	Sequence	
10	CFi9-T7	TAATACGACTCACTATAGGGAGatgacctaataatgatgggttt	(SEQ. ID. NO:20)
10	CFi10c-T7	TAATACGACTCACTATAGGGAGtagtgtgaagggttcatatgc	(SEQ. ID. NO:21)
10	CFi10c-T3	CTCGGAATTAACCCTCACTAAAGGtagtgtgaagggttcatatg	(SEQ. ID. NO:22)
10, 11	CFi10-T7	TAATACGACTCACTATAGGGAGagcatactaaaagtgactctc	(SEQ. ID. NO:23)
11	CFi11c-T7	TAATACGACTCACTATAGGGAGacatgaatgacatttacagcaa	(SEQ. ID. NO:24)
11	CFi11c-T3	CGGAATTAACCCTCACTAAAGGacatgaatgacatttacagcaa	(SEQ. ID. NO:25)

These primers can be used to amplify exon 10 or exon 11 sequences; in another embodiment, multiplex PCR is employed, using two or more pairs of primers to amplify more than one exon at a time.

The product of amplification was then used as a template for the RNA polymerase, with fluoresceinated UTP present to label the RNA product. After sufficient RNA was made, it was fragmented and applied to an exon 10 DNA chip for 15 minutes, after which the chip was washed with hybridization buffer and scanned with the fluorescence microscope. A useful positive control included on many CF exon 10 chips is the 8-mer 3'-CGCCGCCG-5'. FIG. 7, in panels A and B, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to nucleic acid derived from the genomic DNA of an individual with wild-type ΔF508 sequences; in panel B, the target nucleic acid originated from a heterozygous (with respect to the ΔF508 mutation) individual. FIG. 8, in sheets 1 and 2, corresponding to panels A and B of FIG. 7, shows graphs of fluorescence intensity versus tiling position.

These figures show that the sequence of the wild-type RNA can be called for most of the bases near the mutation. In the case of the ΔF508 heterozygous carrier, one particular probe, the same one that distinguished so clearly between the wild-type and mutant oligonucleotide targets in the model system described above, in the T-lane binds a large amount of RNA, while the same probe binds little RNA from the wild-type individual. These results show that the DNA chips of the invention are capable of detecting the ΔF508 mutation in a heterozygous carrier.

Thus, the present invention provides methods for synthesizing large numbers of oligonucleotide probes on a glass substrate and unique probe sets in a defined array in which the probes are arranged in the array by the "tiling" method of the invention. The DNA chips produced by the method can be used to detect mutations in particular sequences of a target nucleic acid, such as genomic DNA or RNA produced from transcription of an amplified genomic DNA. These chips can be used to detect both point mutations and small deletions. Moreover, the pattern of hybridization to the chip allows inferences to be drawn about the sequences of the mutant DNAs.

For example, in the model system involving the cystic fibrosis point mutation G480C, the A-lane probe whose position of substitution corresponds to the position of the mutation does not bind much wild-type target, because in the wild-type sequence, a G occupies that position. However, it binds mutant target very well, allowing one to infer correctly that the mutation involves a change of that G to a T. Similarly, in the case of the three-base deletion in cystic fibrosis known as ΔF508, the T-lane probe that binds mutant target so intensely is responding to the fact that the deletion has brought a CAT sequence into the position occupied by a CTT sequence in the wild-type target. The DNA chips of the invention can be used to detect and sequence not only known mutations in an organism's genome but also new mutations not previously characterized. The DNA chips and methods of the invention can also be used to detect specific sequences in other CFTR exons as well as other human genes for purposes of research and clinical genetic analysis, as demonstrated below.

Detection of Specific Human Mitochondrial DNA Sequences with DNA Chips

As noted above, the present invention provides DNA chips on which a known DNA sequence is represented as an array of overlapping oligonucleotides on a solid support. This set of oligonucleotides is used to probe a target nucleic acid comprising the known sequence, allowing mutations to be detected. As also noted above, there are advantages in some applications to using a minimal set of oligonucleotides specific to the sequence of interest, rather than a set of all possible N-mers. Some of these advantages include: (i) each position in the array is highly informative, whether or not hybridization occurs; (ii) nonspecific hybridization is minimized; (iii) it is straightforward to correlate hybridization differences with sequence differences, particularly with reference to the hybridization pattern of a known standard; and (iv) the ability to address each probe independently during synthesis, using high resolution photolithography, allows the array to be designed and optimized for any sequence. For example the length of any probe can be varied independently of the others.

The present invention illustrates these advantages by providing DNA chips and analytical methods for detecting specific sequences of human mitochondrial DNA. In one preferred embodiment, the invention provides a DNA chip for analyzing sequences contained in a 1.3 kb fragment of human mitochondrial DNA from the "D-loop" region, the most polymorphic region of human mitochondrial DNA. One such chip comprises a set of 269 overlapping oligonucleotide probes of varying length in the range of 9→14 nucleotides with varying overlaps arranged in ~600×600 micron features or synthesis sites in an array 1 cm×1 cm in size. The probes on the chip are shown in columnar form below. An illustrative mitochondrial DNA chip of the invention comprises the following probes (X, Y coordinates are shown, followed by the sequence; "DL3" represents the 3'-end of the probe, which is covalently attached to the chip surface.)

0	0	DL3AGTGGGGTATTT	(SEQ ID. NO:26)
1	0	DL3GGGTATTTAGTT	(SEQ ID. NO:27)
2	0	DL3TTAGTTTATCCAA	(SEQ ID. NO:28)
3	0	DL3ATCCAAACCAGG	(SEQ ID. NO:29)
4	0	DL3ACCAGGATCGGA	(SEQ ID. NO:30)
5	0	DL3CGTGTGTGTGTGG	(SEQ ID. NO:31)
6	0	DL3CGTGTGTGTGTGGC	(SEQ ID. NO:32)
7	0	DL3TCGTGTGTGTGTGG	(SEQ ID. NO:33)
8	0	DL3GTAGGATGGGTC	(SEQ ID. NO:34)
9	0	DL3AGGATGGGTCGT	(SEQ ID. NO:35)
10	0	DL3GATGGGTCGTGT	(SEQ ID. NO:36)
11	0	DL3TGGCGACGATTG	(SEQ ID. NO:37)
12	0	DL3GCGACGATTGGG	(SEQ ID. NO:38)
13	0	DL3TGGGGGGGA	
14	0	DL3GAGGGGGCG	
15	0	DL3GGAGGGGGCGA	(SEQ ID. NO:39)
16	0	DL3GAGGGGGCGA	(SEQ ID. NO:40)
0	1	DL3GGCTTGGTTGG	(SEQ ID. NO:41)
1	1	DL3GGTTGGTTTGGG	(SEQ ID. NO:42)
2	1	DL3TGGGGTTTCTAG	(SEQ ID. NO:43)
3	1	DL3GTTTCTAGTGGG	(SEQ ID. NO:44)
4	1	DL3AGTGGGGGGTGT	(SEQ ID. NO:45)
5	1	DL3GGGGTGTCAAAT	(SEQ ID. NO:46)
6	1	DL3GTCAAATACATCG	(SEQ ID. NO:47)
7	1	DL3ACATCGAATGGAG	(SEQ ID. NO:48)
8	1	DL3CGAATGGAGGAG	(SEQ ID. NO:49)
9	1	DL3GAGGAGTTTCGT	(SEQ ID. NO:50)
10	1	DL3TTTCGTTATGTGA	(SEQ ID. NO:51)
11	1	DL3ATGTGACTTTTAC	(SEQ ID. NO:52)
12	1	DL3GACTTTTACAAAT	(SEQ ID. NO:53)
13	1	DL3AAATCTGCCCGA	(SEQ ID. NO:54)
14	1	DL3AATCTGCCCGAG	(SEQ ID. NO:55)
15	1	DL3CCCGAGTGTAGT	(SEQ ID. NO:56)
16	1	DL3AGTGTAGTGGGG	(SEQ ID. NO:57)
0	2	DL3GGGAGGGTGAG	(SEQ ID. NO:58)
1	2	DL3GGTGAGGGTATG	(SEQ ID. NO:59)
2	2	DL3GGTATGATGATTAG	(SEQ ID. NO:60)
3	2	DL3GATTAGAGTAAGT	(SEQ ID. NO:61)
4	2	DL3TTAGAGTAAGTTA	(SEQ ID. NO:62)
9	2	DL3GGTAGGATGGGT	(SEQ ID. NO:67)
10	2	DL3GGATGGGTCGTG	(SEQ ID. NO:68)
11	2	DL3GGTCGTGTGTGT	(SEQ ID. NO:69)
12	2	DL3GTGTGTGTGGCG	(SEQ ID. NO:70)
13	2	DL3TGTGGCGACGAT	(SEQ ID. NO:71)
14	2	DL3GACGATTGGGGT	(SEQ ID. NO:72)
15	2	DL3ATTGGGGTATGG	(SEQ ID. NO:73)
16	2	DL3GTATGGGGCTTG	(SEQ ID. NO:74)
0	3	DL3GGATTGTGGTCG	(SEQ ID. NO:75)
1	3	DL3TGGTCGGATTGG	(SEQ ID. NO:76)
2	3	DL3GGATTGGTCTAAA	(SEQ ID. NO:77)
3	3	DL3TCTAAAGTTTAAA	(SEQ ID. NO:78)
4	3	DL3GTTTAAAATAGAA	(SEQ ID. NO:79)
5	3	DL3ATAGAAAAACCG	(SEQ ID. NO:80)
6	3	DL3AGAAAAACCGC	(SEQ ID. NO:81)
7	3	DL3AACCGCCATAC	(SEQ ID. NO:82)
8	3	DL3CCATACGTGAAAA	(SEQ ID. NO:83)
9	3	DL3ACGTGAAAATTGT	(SEQ ID. NO:84)
10	3	DL3AATTGTCAGTGGG	(SEQ ID. NO:85)
11	3	DL3TGTCAGTGGGGG	(SEQ ID. NO:86)
12	3	DL3TGGGGTTGA	(SEQ ID. NO:87)
13	3	DL3GGGTTGATTGTGT	(SEQ ID. NO:88)
14	3	DL3TTGTGTAATAAAA	(SEQ ID. NO:89)
15	3	DL3AATAAAAGGGGA	(SEQ ID. NO:90)
16	3	DL3TAAAAGGGGAGG	(SEQ ID. NO:91)
0	4	DL3GTTTTTTAAAGG	(SEQ ID. NO:92)
1	4	DL3TTTTAAAGGTGG	(SEQ ID. NO:93)
2	4	DL3AGGTGGTTTGG	(SEQ ID. NO:94)
3	4	DL3TTGGGGGGGAG	(SEQ ID. NO:95)
4	4	DL3GGAGGGGGCG	(SEQ ID. NO:96)
5	4	DL3GGGGCGAAGAC	(SEQ ID. NO:97)
6	4	DL3GAAGACCGGATG	(SEQ ID. NO:98)
7	4	DL3CCGGATGTCGTG	(SEQ ID. NO:99)
8	4	DL3GTCGTGAATTTGT	(SEQ ID. NO:100)
9	4	DL3CGTGAATTTGTGT	(SEQ ID. NO:101)
10	4	DL3TTGTGTAGAGACG	(SEQ ID. NO:102)
11	4	DL3TAGAGACGGTTT	(SEQ ID. NO:103)
12	4	DL3ACGGTTTGGGG	(SEQ ID. NO:104)
13	4	DL3TGGGGTTTTTGT	(SEQ ID. NO:105)
14	4	DL3GGGTTTTTGTTT	(SEQ ID. NO:106)

-continued

5	2	DL3AAGTTATGTTGGG	(SEQ ID. NO:63)	15	4	DL3TTGTTTCTTGGG	(SEQ ID. NO:107)
6	2	DL3GTTGGGGGCG	(SEQ ID. NO:64)	16	4	DL3TCTTGGGATTGTG	(SEQ ID. NO:108)
7	2	DL3GGGGGGGGTA	(SEQ ID. NO:65)	0	5	DL3TGTATGAATGATTT	(SEQ ID. NO:109)
8	2	DL3GCGGGTAGGAT	(SEQ ID. NO:66)	1	5	DL3TGATTTCACACAA	(SEQ ID. NO:110)
2	5	DL3ACACAATTAATTAA	(SEQ ID. NO:111)	14	7	DL3CTCTGCGACCTC	(SEQ ID. NO:157)
3	5	DL3AATTAATTACGAA	(SEQ ID. NO:112)	15	7	DL3GACCTCGGCCT	(SEQ ID. NO:158)
4	5	DL3TACGAACATCCTG	(SEQ ID. NO:113)	16	7	DL3TCGGCCTCGTG	(SEQ ID. NO:159)
5	5	DL3ACGAACATCCTGT	(SEQ ID. NO:114)	0	8	DL3GATGAAGTCCCAG	(SEQ ID. NO:160)
6	5	DL3TCCTGTATTATTA	(SEQ ID. NO:115)	1	8	DL3AGTCCCAGTATTT	(SEQ ID. NO:161)
7	5	DL3GTATTATTATTGTT	(SEQ ID. NO:116)	2	8	DL3GTATTTCGGATTT	(SEQ ID. NO:162)
8	5	DL3ATTGTTAAACTTA	(SEQ ID. NO:117)	3	8	DL3TCGGATTTATCG	(SEQ ID. NO:163)
9	5	DL3AAACTTACAGACG	(SEQ ID. NO:118)	4	8	DL3GATTTATCGGGT	(SEQ ID. NO:164)
10	5	DL3ACAGACGTGTCG	(SEQ ID. NO:119)	5	8	DL3ATCGGGTGTGCA	(SEQ ID. NO:165)
11	5	DL3GTGTCGGTGAAA	(SEQ ID. NO:120)	6	8	DL3TGTGCAAGGGGA	(SEQ ID. NO:166)
12	5	DL3GTGAAAGGTGTGT	(SEQ ID. NO:121)	7	8	DL3CAAGGGGAATTT	(SEQ ID. NO:167)
13	5	DL3GGTGTGTCTGTAG	(SEQ ID. NO:122)	8	8	DL3GAATTTATTCTGTA	(SEQ ID. NO:168)
14	5	DL3TGTGTCTGTAGTA	(SEQ ID. NO:123)	9	8	DL3TCTGTAGTGCTAC	(SEQ ID. NO:169)
15	5	DL3GTAGTATTGTTTT	(SEQ ID. NO:124)	10	8	DL3GTAGTGCTACCT	(SEQ ID. NO:170)
16	5	DL3AGTATTGTTTTTT	(SEQ ID. NO:125)	11	8	DL3GCTACCTAGTAG	(SEQ ID. NO:171)
0	6	DL3CCTCGTGGGATA	(SEQ ID. NO:126)	12	8	DL3CTAGTAGTCCAGA	(SEQ ID. NO:172)
1	6	DL3TGGGATACAGCG	(SEQ ID. NO:127)	13	8	DL3TCCAGATA9TGGG	(SEQ ID. NO:173)
2	6	DL3GATACAGCGTCAT	(SEQ ID. NO:128)	14	8	DL3AGATAGTGGGATA	(SEQ ID. NO:174)
3	6	DL3GCGTCATAGACAG	(SEQ ID. NO:129)	15	8	DL3GGGATAATTGGT	(SEQ ID. NO:175)
4	6	DL3AGACAGAAACTAA	(SEQ ID. NO:130)	16	8	DL3TAATTGGTGAGTG	(SEQ ID. NO:176)
5	6	DL3CAGAAACTAAGGA	(SEQ ID. NO:131)	0	9	DL3TATAGGGCGTGT	(SEQ ID. NO:177)
6	6	DL3TAAGGACGGAGT	(SEQ ID. NO:132)	1	9	DL3GGGCGTGTTCTCA	(SEQ ID. NO:178)
7	6	DL3GACGGAGTAGGA	(SEQ ID. NO:133)	2	9	DL3GTGTTCTCACGAT	(SEQ ID. NO:179)
8	6	DL3GTAGGATAATAAA	(SEQ ID. NO:134)	3	9	DL3TCACGATGAGAGG	(SEQ ID. NO:180)
9	6	DL3TAATAAATAGCG	(SEQ ID. NO:135)	4	9	DL3ATGAGAGGAGCG	(SEQ ID. NO:181)
10	6	DL3ATAGCGTAGGAT	(SEQ ID. NO:136)	5	9	DL3AGGAGCGAGGC	(SEQ ID. NO:182)
11	6	DL3TAGCGTAGGATG	(SEQ ID. NO:137)	6	9	DL3CGAGGCCGG	(SEQ ID. NO:183)
12	6	DL3AGGATGCAAGTT	(SEQ ID. NO:138)	7	9	DL3GCCCGGGTATT	(SEQ ID. NO:184)
13	6	DL3ATGCAAGTTATAA	(SEQ ID. NO:139)	8	9	DL3CGGGTATTGTGA	(SEQ ID. NO:185)
14	6	DL3GTTATAATGTCCG	(SEQ ID. NO:140)	9	9	DL3GTGAACCCCCAT	(SEQ ID. NO:186)
15	6	DL3ATGTCCGCTTGT	(SEQ ID. NO:141)	10	9	DL3CCCCATCGATTT	(SEQ ID. NO:187)
16	6	DL3TCCGCTTGTATG	(SEQ ID. NO:142)	11	9	DL3ATCGATTTCACTT	(SEQ ID. NO:188)
0	7	DL3GTGAGTGCCCTC	(SEQ ID. NO:143)	12	9	DL3TTTCACTTGACAT	(SEQ ID. NO:189)
1	7	DL3TGCCCTCGAGAG	(SEQ ID. NO:144)	13	9	DL3TTGACATAGAGCT	(SEQ ID. NO:190)
2	7	DL3CCTCGAGAGGTA	(SEQ ID. NO:145)	14	9	DL3TAGAGCTGTAGAC	(SEQ ID. NO:191)
3	7	DL3AGAGGTACGTAA	(SEQ ID. NO:146)	15	9	DL3GTAGACCAAGGA	(SEQ ID. NO:192)
4	7	DL3ACGTAAACCATA	(SEQ ID. NO:147)	16	9	DL3ACCAAGGATGAAG	(SEQ ID. NO:193)
5	7	DL3ACCATAAAAGCAG	(SEQ ID. NO:148)	0	10	DL3CGTGTAATGTCAG	(SEQ ID. NO:194)
6	7	DL3AAAGCAGACCC	(SEQ ID. NO:149)	1	10	DL3TGTCAGTTTAGGG	(SEQ ID. NO:195)
7	7	DL3AGACCCCCCAT	(SEQ ID. NO:150)	2	10	DL3TCAGTTTAGGGA	(SEQ ID. NO:196)
8	7	DL3CCCCCATACGT	(SEQ ID. NO:151)	3	10	DL3TAGGGAAGAGCA	(SEQ ID. NO:197)
9	7	DL3CATACGTGCGCT	(SEQ ID. NO:152)	4	10	DL3AAGAGCAGGGGT	(SEQ ID. NO:198)
10	7	DL3GTGCGCTATCAG	(SEQ ID. NO:153)	5	10	DL3CAGGGGTACCTA	(SEQ ID. NO:199)
11	7	DL3GCGCTATCAGTA	(SEQ ID. NO:154)	6	10	DL3GGTACCTACTGG	(SEQ ID. NO:200)
12	7	DL3TCAGTAACGCTC	(SEQ ID. NO:155)	7	10	DL3TACTGGGGGGA	(SEQ ID. NO:201)
13	7	DL3GTAACGCTCTGC	(SEQ ID. NO:156)	8	10	DL3GGGGGAGTCTAT	(SEQ ID. NO:202)
9	10	DL3AGTCTATCCCCA	(SEQ ID. NO:203)	11	13	DL3CATGTATTTTTGG	(SEQ ID. NO:246)
10	10	DL3ATCCCCAGGGA	(SEQ ID. NO:204)	12	13	DL3TTTTGGGTTAGG	(SEQ ID. NO:247)
11	10	DL3CAGGGAACTGGT	(SEQ ID. NO:205)	13	13	DL3GGGTTAGGATGT	(SEQ ID. NO:248)
12	10	DL3ACTGGTGGTAGG	(SEQ ID. NO:206)	14	13	DL3GGATGTAGTTTTG	(SEQ ID. NO:249)
13	10	DL3CTGGTGGTAGGA	(SEQ ID. NO:207)	15	13	DL3TGTAGTTTTGGG	(SEQ ID. NO:250)
14	10	DL3GTAGGAGGCACA	(SEQ ID. NO:208)	16	13	DL3TTTGGGGGAGG	(SEQ ID. NO:251)
15	10	DL3GGCACATTTAGT	(SEQ ID. NO:209)	5	14	DL3GGGTTCATAACTG	(SEQ ID. NO:252)
16	10	DL3TTTAGTTATAGGG	(SEQ ID. NO:210)	6	14	DL3ATAACTGAGTGGG	(SEQ ID. NO:253)
0	11	DL3AGGTTTACGGTG	(SEQ ID. NO:211)	7	14	DL3AACTGAGTGGGT	(SEQ ID. NO:254)
1	11	DL3TACGGTGGGGA	(SEQ ID. NO:212)	8	14	DL3GTGGGTAGTTGT	(SEQ ID. NO:255)
2	11	DL3GTGGGGAGTGG	(SEQ ID. NO:213)	9	14	DL3GTAGTTGTTGGC	(SEQ ID. NO:256)
3	11	DL3GGGAGTGGGTGA	(SEQ ID. NO:214)	10	14	DL3GTTGGCGATACA	(SEQ ID. NO:257)
4	11	DL3GGGTGATCCTATG	(SEQ ID. NO:215)	11	14	DL3CGATACATAAAAG	(SEQ ID. NO:258)
5	11	DL3CCTATGGTTGTTT	(SEQ ID. NO:216)	12	14	DL3TAAAAGCATGTAA	(SEQ ID. NO:259)
6	11	DL3GGTTGTTTGGATG	(SEQ ID. NO:217)	13	14	DL3GCATGTAATGACG	(SEQ ID. NO:260)
7	11	DL3GTTTGGATGGGT	(SEQ ID. NO:218)	14	14	DL3ATGACGGTCGGT	(SEQ ID. NO:261)
8	11	DL3ATGGGTGGGAAT	(SEQ ID. NO:219)	15	14	DL3GTCGGTGGTACT	(SEQ ID. NO:262)
9	11	DL3GGGAATTGTCATG	(SEQ ID. NO:220)	16	14	DL3GGTACTTATAACA	(SEQ ID. NO:263)
10	11	DL3GTCATGTATCATGT	(SEQ ID. NO:221)	5	15	DL3TCGATTCTAAGAT	(SEQ ID. NO:264)
11	11	DL3TCATGTATTTCGG	(SEQ ID. NO:222)	6	15	DL3TAAGATTAAATTT	(SEQ ID. NO:265)
12	11	DL3TATTTCGGTAAA	(SEQ ID. NO:223)	7	15	DL3AAATTTGAATAAG	(SEQ ID. NO:266)
13	11	DL3TTCGGTAAATGG	(SEQ ID. NO:224)	8	15	DL3AATAAGAGACAAG	(SEQ ID. NO:267)
14	11	DL3GTAAATGGCATGT	(SEQ ID. NO:225)	9	15	DL3AAGAGACAAGAAA	(SEQ ID. NO:268)
15	11	DL3GCATGTAATCGTG	(SEQ ID. NO:226)	10	15	DL3AAGAAAGTACCC	(SEQ ID. NO:269)
16	11	DL3GTAATCGTGTAAT	(SEQ ID. NO:227)	11	15	DL3AAAGTACCCCTT	(SEQ ID. NO:270)
5	12	DL3GGGAGGGGTAC	(SEQ ID. NO:228)	12	15	DL3CCCCTTCGTCTA	(SEQ ID. NO:271)
6	12	DL3GGGTACGAATGT	(SEQ ID. NO:229)	13	15	DL3CTTCGTCTAAAC	(SEQ ID. NO:272)
7	12	DL3ACGAATGTTCGTT	(SEQ ID. NO:230)	14	15	DL3CTAAACCCATGG	(SEQ ID. NO:273)
8	12	DL3TGTTCGTTCATGT	(SEQ ID. NO:231)	15	15	DL3AACCCATGGTGG	(SEQ ID. NO:274)
9	12	DL3CGTTCATGTCGTT	(SEQ ID. NO:232)	16	15	DL3TGGTGGGTTCAT	(SEQ ID. NO:275)

-continued

10	12	DL3GTCGTTAGTTGG	(SEQ ID. NO:233)	5	16	DL3TTGGAAAAAGGT	(SEQ ID. NO:276)
11	12	DL3TAGTTGGGAGTT	(SEQ ID. NO:234)	6	16	DL3AAAAGGTTCCTG	(SEQ ID. NO:277)
12	12	DL3GGAGTTGATAGTG	(SEQ ID. NO:235)	7	16	DL3GGTTCCTGTTTA	(SEQ ID. NO:278)
13	12	DL3ATAGTGTGTAGTT	(SEQ ID. NO:236)	8	16	DL3CCTGTTTAGTCTC	(SEQ ID. NO:279)
14	12	DL3GTFTAGTTGACGT	(SEQ ID. NO:237)	9	16	DL3TTAGTCTCTTTTT	(SEQ ID. NO:280)
15	12	DL3TGACGTTGAGGT	(SEQ ID. NO:238)	10	16	DL3CTTTTTCAGAAAT	(SEQ ID. NO:281)
16	12	DL3CGTTGAGGTTTA	(SEQ ID. NO:239)	11	16	DL3AGAAATTGAGGTG	(SEQ ID. NO:282)
5	13	DL3TATAACATGCCAT	(SEQ ID. NO:240)	12	16	DL3AAATTGAGGTGGT	(SEQ ID. NO:283)
6	13	DL3AACATGCCATGGT	(SEQ ID. NO:241)	13	16	DL3GGTGGTAATCGT	(SEQ ID. NO:284)
7	13	DL3CCATGGTATTAT	(SEQ ID. NO:242)	14	16	DL3TAATCGTGGGTT	(SEQ ID. NO:285)
8	13	DL3ATTTATGAACTGG	(SEQ ID. NO:243)	15	16	DL3GTGGGTTTCGAT	(SEQ ID. NO:286)
9	13	DL3AACTGGTGGACAT	(SEQ ID. NO:244)	16	16	DL3GGTTTCGATTCT	(SEQ ID. NO:287)
10	13	DL3TGGACATCATGTA	(SEQ ID. NO:245)				

No probes were present in positions X, Y=0, 12 to X, Y=4, 12; X, Y=0, 13 to X, Y=4, 13; X, Y=0, 14 to X, Y=4, 14; X, Y=0, 15 to X, Y=4, 15; X, Y=0, 16 to X, Y=4, 16; The length of each of the probes on the chip was variable to minimize differences in melting temperature and potential for cross-hybridization. Each position in the sequence is represented by at least one probe and most positions are represented by 2 or more probes. As noted above, the amount of overlap between the oligonucleotides varies from probe to probe. FIG. 9 shows the human mitochondrial genome; "O_H" is the H strand origin of replication, and arrows indicate the cloned unshaded sequence.

DNA was prepared from hair roots of six human donors (mt1 to mt6) and then amplified by PCR and cloned into M13; the resulting clones were sequenced using chain terminators to verify that the desired specific sequences were present. DNA from the sequenced M13 clones was amplified by PCR, transcribed in vitro, and labeled with fluorescein-UTP using T3 RNA polymerase. The 1.3 kb RNA transcripts were fragmented and hybridized to the chip. The results showed that each different individual had DNA that produced a unique hybridization fingerprint on the chip and that the differences in the observed patterns could be correlated with differences in the cloned genomic DNA sequence. The results also demonstrated that very long sequences of a target nucleic acid can be represented comprehensively as a specific set of overlapping oligonucleotides and that arrays of such probe sets can be usefully applied to genetic analysis.

The sample nucleic acid was hybridized to the chip in a solution composed of 6xSSPE, 0.1% Triton-X 100 for 60 minutes at 15° C. The chip was then scanned by confocal scanning fluorescence microscopy. The individual features on the chip were 588x588 microns, but the lower left 5x5 square features in the array did not contain probes. To quantitate the data, pixel counts were measured within each synthesis site. Pixels represent 50x50 microns. The fluorescence intensity for each feature was scaled to a mean determined from 27 bright features. After scanning, the chip was stripped and rehybridized; all six samples were hybridized to the same chip. FIG. 10 shows the image observed from the mt4 sample on the DNA chip. FIG. 11 shows the image observed from the mt5 sample on the DNA chip. FIG. 12 shows the predicted difference image between the mt4 and mt5 samples on the DNA chip based on mismatches between the two samples and the reference sequence (see Anderson et al., 1981, *Nature* 290: 457–465, incorporated herein by reference). FIG. 13 shows the actual difference image observed.

The results show that, in almost all cases, mismatched probe/target hybrids resulted in lower fluorescence intensity than perfectly matched hybrids. Nonetheless, some probes detected mutations (or specific sequences) better than others, and in several cases, the differences were within noise levels. Improvements can be realized by increasing the amount of overlap between probes and hence overall probe density and, for duplex DNA targets, using a second set of probes, either on the same or a separate chip, corresponding to the second strand of the target. FIG. 14, in sheets 1 and 2, shows a plot of normalized intensities across rows 10 and 11 of the array and a tabulation of the mutations detected.

FIG. 15 shows the discrimination between wild-type and mutant hybrids obtained with this chip. The median of the six normalized hybridization scores for each probe was taken. The graph plots the ratio of the median score to the normalized hybridization score versus mean counts. On this graph, a ratio of 1.6 and mean counts above 50 yield no false positives, and while it is clear that detection of some mutants can be improved, excellent discrimination is achieved, considering the small size of the array. FIG. 16 illustrates how the identity of the base mismatch may influence the ability to discriminate mutant and wild-type sequences more than the position of the mismatch within an oligonucleotide probe. The mismatch position is expressed as % of probe length from the 3'-end. The base change is indicated on the graph. These results show that the DNA chip increases the capacity of the standard reverse dot blot format by orders of magnitude, extending the power of that approach many fold and that the methods of the invention are more efficient and easier to automate than gel-based methods of nucleic acid sequence and mutation analysis.

These advantages become more apparent as chips with more and more probes are employed. To illustrate, the present invention provides a DNA chip for analyzing human mitochondrial DNA (mtDNA) that "tiles" through 648 nucleotides of human H strand mtDNA from positions 16280 to 356. The probes in the array are 15 nucleotides in length, and each position in the target sequence is represented by a set of 4 probes (A, C, G, T substitutions), which differed from one another at position 7 from the 3'-end. The array consists of 13 blocks of 4x50 probes: each block scans through 50 nucleotides of contiguous mtDNA sequence. The blocks are separated by blank rows. The 4 corner columns contain control probes; there are a total of 2600 probes in a 1.28 cmx1.28 cm square area (feature), and each area is 256x197 microns.

Labeled RNA target DNA was prepared by PCR amplification of a 1.3 kb region of human mtDNA spanning positions 15935 to 667, cloning into M13 (sequence verification was performed), and reamplification of the cloned sequences using primers tagged with T3 and T7 RNA polymerase promoter sequences and in vitro transcription to produce fluorescein-UTP labeled RNA. The RNA was fragmented and hybridized to the oligonucleotide array in a solution composed of 6xSSPE, 0.1% Triton X-100 for 60 minutes at 18° C. Unhybridized material was washed away with buffer, and the chip was scanned at 25 micron pixel resolution.

FIG. 17 provides a 5' to 3' sequence listing of one target corresponding to the probes on the chip. X is a control probe. Positions that differ in the target (i.e., are mismatched with the probe at the designated site) are in bold. FIG. 18 shows the fluorescence image produced by scanning the chip when hybridized to this sample. About 95% of the sequence could be read correctly from only one strand of the original duplex target nucleic acid. Although some probes did not provide excellent discrimination and some probes did not appear to hybridize to the target efficiently, excellent results were achieved. The target sequence differed from the probe set at six positions: 4 transitions and 2 insertions. All 4 transitions were detected, and specific probes could readily be incorporated into the array to detect insertions or deletions. FIG. 19 illustrates the detection of 4 transitions in the target sequence relative to the wild-type probes on the chip.

These results illustrate that longer sequences can be read using the DNA chips and methods of the invention, as compared to conventional sequencing methods, where reading length is limited by the resolution of gel electrophoresis. Similar results were observed when genomic DNA samples were prepared from human hair roots. Hybridization and signal detection require less than an hour and can be readily shortened by appropriate choice of buffers, temperatures, probes, and reagents. In principle, longer sequence reads can be obtained than by conventional sequencing, where reading length is limited by the resolution of gel electrophoresis.

P53 Sequencing and Diagnostic DNA Chips

P53 is a tumor suppressor gene that has been found to be mutated in most forms of cancer (see Levine et al, 1991, *Nature* 351: 453–456, and Hollstein et al., 1991, *Science* 253: 49–53, each of which is incorporated herein by reference). In addition, there is a hereditary syndrome, Li-Fraumeni, in which individuals inherit mutant alleles of p53 and tend to have cancer at relatively young ages (Frebourg et al., 1992, *PNAS* 89: 6413–6417, incorporated herein by reference). During the development of a cancer, p53 is inactivated. The course of p53 inactivation generally involves a mutation in one copy of p53 and is often followed by deletion of the other copy. After p53 is inactivated, chromosomal abnormalities begin to appear in tumors. In the best understood form of cancer, colorectal cancer, well over 50%, perhaps 80%, of all patients with tumors have p53 mutations. In addition, p53 mutations have been found in a high proportion of lung, breast, and other tumors (Rodrigues et al., 1990, *PNAS* 87: 7555–7559, incorporated herein by reference). According to data presented by David Sidransky (1992 San Diego Conference), over 400 mutations in p53 are known.

The p53 gene spans 20 kbp in humans and has 11 exons, 10 of which are protein coding (see Tominaga et al., 1992, *Critical Reviews in Oncogenesis* 3: 257–282, incorporated herein by reference). The gene produces a 53 kilodalton phosphoprotein that regulates DNA replication. The protein acts to halt replication at the G1/S boundary in the cell cycle and is believed to act as a "molecular policeman," shutting down replication when the DNA is damaged or blocking the reproduction of DNA viruses (see Lane, 1992, *Nature* 358: 15–16, incorporated herein by reference). There is substantial interest in the cancer research community in analyzing p53 mutations. The NCI is currently funding contracts to characterize the p53 mutation spectra caused by various carcinogens. In addition, there are research projects which involve sequencing p53 from spontaneously arising tumors. A major resource in these studies is the huge supply of biopsy material stored in paraffin blocks. Also, there are projects which are aimed at analyzing the relationship between the particular mutation in p53 and the functioning of the resulting protein. Furthermore, there are projects looking at the germline inheritance of p53 mutations and the development of cancer. The present invention provides useful DNA chips and methods for such studies.

In addition, the present invention also provides a diagnostic test kit and method and p53 probes immobilized on a DNA chip in an organized array. Currently available diagnostic tests for cancer typically have a sensitivity of about 50%. The present invention provides significant advantages over such tests, and in one embodiment provides a method for detecting cancer-causing mutations in p53 that involves the steps of (1) obtaining a biopsy, which is optionally fractionated by cryostat sectioning to enrich tumor cells to about 80% of the total cell population. The DNA or RNA is then extracted, amplified, and analyzed with a DNA chip for the presence of p53 mutations correlated with malignancy.

To illustrate the value of the DNA chips of the present invention in such a method, a DNA chip was synthesized by the VLSIPS™ method to provide an array of overlapping probes which represent or tile across a 60 base region of exon 6 of the p53 gene. To demonstrate the ability to detect substitution mutations in the target, twelve different single substitution mutations (wild type and three different substitutions at each of three positions) were represented on the chip along with the wild type. Each of these mutations was represented by a series of twelve 12-mer oligonucleotide probes, which were complementary to the wild type target except at the one substituted base. Each of the twelve probes was complementary to a different region of the target and contained the mutated base at a different position, e.g., if the substitution was at base 32, the set of probes would be complementary–with the exception of base 32—to regions of the target 21–32, 22–33, and 32–43). This enabled investigation of the effect of the substitution position within the probe. The alignment of some of the probes with a 12-mer model target nucleic acid is shown in FIG. 20.

To demonstrate the effect of probe length, an additional series of ten 10-mer probes was included for each mutation (see FIG. 21). In the vicinity of the substituted positions, the wild-type sequence was represented by every possible overlapping 12-mer and 10-mer probe. To simplify comparisons, the probes corresponding to each varied position were arranged on the chip in the rectangular regions with the following structure: each row of cells represents one substitution, with the top row representing the wild type. Each column contains probes complementary to the same region of the target, with probes complementary to the 3'-end of the target on the left and probes complementary to the 5'-end of the target on the right. The difference between two adjacent columns is a single base shift in the positioning of the probes. Whenever possible, the series of 10-mer probes were placed in four rows immediately underneath and aligned with the 4 rows of 12-mer probes for the same mutation.

To provide model targets, 5' fluoresceinated 12-mers containing all possible substitutions in the first position of codon 192 were synthesized (see the starred position in the target in FIG. 20). Solutions containing 10 nM target DNA in 6xSSPE, 0.25% Triton X-100 were hybridized to the chip at room temperature for several hours. While target nucleic was hybridized to the chip, the fluorophores on the chip were excited by light from an argon laser, and the chip was scanned with an autofocusing confocal microscope. The emitted signals were processed by a PC to produce an image using image analysis software. By 1 to 3 hours, the signal had reached a plateau; to remove the hybridized target and allow hybridization to another target, the chip was stripped with 60% formamide, 2xSSPE at 17° C. for 5 minutes. The washing buffer and temperature can vary, but the buffer typically contains 2-to-3xSSPE, 10-to-60% formamide (one can use multiple washes, increasing the formamide concentration by 10% each wash, and scanning between washes to determine when the wash is complete), and optionally a small percentage of Triton X-100, and the temperature is typically in the range of 15° to 18° C.

Very distinct patterns were observed after hybridization with targets with 1 base substitutions and visualization with a confocal microscope and software analysis, as shown in FIG. 22. In general, the probes which form perfect matches with thê target retain the highest signal. For example, in the first image in Figure PC, the 12-mer probes that form perfect matches with the wild-type (WT) target are in the first row (top). The 12-mer probes with single base mismatches are located in the second, third, and fourth rows and have much lower signals. The data is also depicted graphically in FIG. 23. On each graph, the X ordinate is the position of the probe in its row on the chip, and the Y ordinate is the signal at that probe site after hybridization.

When a target with a different one base substitution is hybridized the complementary set of probes has the highest signal (see pictures 2, 3, and 4 in FIG. 22 and graphs 2, 3, and 4 in FIG. 23). In each case, the probe set with no mismatches with the target has the highest signals. Within a 12-mer probe set, the signal was highest at position 6 or 7. The graphs show that the signal difference between 12-mer probes at the same X ordinate tended to be greatest at positions 5 and 8 when the target and the complementary probes formed 10 base pairs and 11 base pairs, respectively. Because tumors often have both WT and mutant p53 genes, mixed target populations were also hybridized to the chip, as shown in FIG. 24. When the hybridization solution consisted of a 1:1 mixture of WT 12-mer and a 12-mer with a substitution in position 7 of the target, the sets of probes that were perfectly matched to both targets showed higher signals than the other probe sets.

The hybridization efficiency of a 10-mer probe array as compared to a 12-mer probe array was also compared. The 10-mer and 12-mer probe arrays gave comparable signals (see graphs 1–4 in FIG. 23 and graphs 1–4 in FIG. 25). However, the 10-mer probe sets, which are in rows 5–8 (see images in FIG. 22), seemed to be better in this model system than the 12-mer probe sets at resolving one target from another, consistent with the expectation that one base mismatches are more destabilizing for 10-mers than 12-mers. Hybridization results within probe sets perfectly matched to target also followed the expectation that, the more matches the individual probe formed with the target, the higher the signal. However, duplexes with two 3' dangles (see FIG. 23, position 6 in graphs 1–4) have about as much signal as the probes which are matched along their entire length (see FIG. 23, position 7, in graphs 1–4).

This illustrative model system shows that 12-mer targets that differ by one base substitutions can be readily distinguished from one another by the novel probe array provided by the invention and that resolution of the different 12-mer targets was somewhat better with the 10-mer probe sets than with the 12-mer probe sets. The value of having several overlapping probes hybridizing to a target demonstrates the value of the multiple hybridization events that take place on a DNA chip of the invention. The results also demonstrate the feasibility of constructing a probe set to sequence the entire 1.4 kbp protein coding region of p53 or alternatively the 0.6 kbp of exons 5–9 containing mutation hot spots.

For sequencing, the p53 DNA can be cloned from the sample or directly amplified from genomic DNA by PCR. If genomic PCR is used, then the DNA can be diluted prior to amplification so that a single copy of the gene is amplified. For diagnostic purposes, the genomic DNA can be isolated from a tumor biopsy in which the tumor cells may be the majority population. As noted above, the proportion of tumor cells in a sample can be enriched by cryostat sectioning. DNA can also be isolated and amplified from tumor samples stored in paraffin blocks.

The p53 DNA in the sample can be amplified by PCR (although other amplification methods can be used) using 3–4 primer pairs generating amplicons of <3 kbp each. Illustrative primers of the invention for amplifying exon 5 of the p53 gene are shown below (B is biotin; F is fluorescein).

5'-B-CACTTGTGCCCTGACTTTCAAC-3'(SEQ. ID NO:288)

5'-F-CACTTGTGCCCTGACTTTCAAC-3'

5'-ATGCAATTAACCCTCACTAAAGGGAGACACTTGTGCCCTGACTTTCAAC-3'(SEQ. ID NO:289) (has T3 promoter)

5'-B-GACCCTGGGCAACCAGCCCTGTCGT-3'(SEQ. ID NO:290)

5'-F-GACCCTGGGCAACCAGCCCTGTCGT-3'

5'-TAATACGACTCACTATAGGGAGGACCCTGGGCAACCAGCCCTGTCGT-3'(SEQ. ID NO:291) (has T3 promoter)

After PCR amplification of the target (the amplified target is called the "amplicon") one strand of the amplicon can then be isolated, i.e., using a biotinylated primer that allows capture of the undesired strand on streptavidin beads. Alternatively, asymmetric PCR can be used to generate a single-stranded target. Another approach involves the generation of single stranded RNA form the PCR product by incorporating a T7 or other RNA polymerase promoter in one of the primers. The single-stranded material can optionally be fragmented to generate smaller nucleic acids with less significant secondary structure than longer nucleic acids.

In one such method, fragmentation is combined with labeling. To illustrate, degenerate 8-mers or other degenerate short oligonucleotides are hybridized to the single-stranded target material. In the next step, a DNA polymerase is added with the four different dideoxynucleotides, each labeled with a different fluorophore. Fluorophore-labeled dideoxynucleotide are available from a variety of commercial suppliers, such as ABI. Hybridized 8-mers are extended by a labeled dideoxynucleotide. After an optional purification step, i.e., with a size exclusion column, the labeled 9-mers are hybridized to the chip. Other methods of target fragmentation can be employed. The single-stranded DNA can be fragmented by partial degradation with a DNAse or partial depurimation with acid. Labeling can be accomplished in a separate step, i.e., fluorophore-labeled nucleotides are incorporated before the fragmentation step or a DNA binding fluorophore, such as ethidium homodimer, is attached to the target after fragmentation.

In one embodiment, the DNA chip has an array of 10^4 to 10^5 probes tiling across the protein coding regions of p53, which comprise about 1200 bp; smaller arrays specific for the 600 bp mutational hot spot region are also useful. The probes overlap for N-2 to N-4 bases, where N is the length of the probe in bases. N is typically 10 to 14 bases long, but as will be seen below, probes 15 to 19 bases and longer are also useful. Every possible single base substitution occurring one at a time is represented in the array. The number of unique 10-mer probes with 7 base overlaps would be about

(1200/3)×4×10 or about 1.6×10^4. To allow 3 replicates of each probe, one might have a total array size on the order of 4.8×10^4 probes. Of course, arrays of probes within the ranges of 10^2 to 10^6 probes are also useful for applications; for example, very large arrays of 10^6 or more probes are useful for sequencing or sequence checking large genomic DNA fragments. Optionally fragmented and labeled target nucleic acid hybridized to the chip is detected by a confocal microscope or other imaging device. The pattern of sites "lighting up" with target is preferably analyzed with computer assistance to provide the sequence of the target from the pattern of sites producing signals.

The invention is illustrated below with examples of DNA chips comprising very large arrays of DNA probes to "resequence" p53 target nucleic acid in a sample. To analyze DNA from exon 5 of the p53 tumor suppressor gene, a set of overlapping 17-mer probes was synthesized on a chip. The probes for the WT allele were synthesized so as to tile across the entire exon with single base overlaps between probes. For each WT probe, a sets of 4 additional probes, one for each possible base substitution at position 7, were synthesized and placed in a column relative to the WT probe. Exon 5 DNA was amplified by PCR with primers flanking the exon. One of the primers was labeled with fluorescein; the other primer was labeled with biotin. After amplification, the biotinylated strand was removed by binding to streptavidin beads. The fluoresceinated strand was used in hybridization.

About ⅓ of the amplified, single-stranded nucleic acid was hybridized overnight in 5xSSPE at 60° C. to the probe chip (under a cover slip). After washing with 6xSSPE, the chip was scanned using confocal microscopy. FIG. 26 shows an image of the p53 chip hybridized to the target DNA. Analysis of the intensity data showed that 93.5% of the 184 bases of exon 5 were called in agreement with the WT sequence (see Buchman et al., 1988, *Gene* 70: 245–252, incorporated herein by reference). The miscalled bases were from positions where probe signal intensities were tied (1.6%) and where non-WT probes had the highest signal intensity (4.9%). FIG. 27 illustrates how the actual sequence was read. Gaps in the sequence of letters in the WT rows correspond to control probes or sites. Positions at which bases are miscalled are represented by letters in italic type in cells corresponding to probes in which the WT bases have been substituted by other bases.

As the diagram indicates, the miscalled bases are from the low intensity areas of the image, which may be due to secondary structure in the target or probes preventing intermolecular hybridization. To diminish the effects due to secondary structure, one can employ shorter targets (i.e., by target fragmentation) or use more stringent hybridization conditions. In addition, the use of a set of probes synthesized by tiling across the other strand of a duplex target can also provide sequence information buried in secondary structure in the other strand. It should be appreciated, however, that the pattern of low intensity areas that forms as a result of secondary structure in the target itself provides a means to identify that a specific target sequence is present in a sample. Other factors that may contribute to lower signal intensities include differences in probe densities and hybridization stabilities.

These results demonstrate the advantages provided by the DNA chips of the invention to genetic analysis. As another example, heterozygous mutations are currently sequenced by an arduous process involving cloning and repurification of DNA. The cloning step is required, because the gel sequencing systems are poor at resolving even a 1:1 mixture of DNA. First, the target DNA is amplified by PCR with primers allowing easy ligation into a vector, which is taken up by transformation of *E. coli* which in turn must be cultured, typically on plates overnight. After growth of the bacteria, DNA is purified in a procedure that typically takes about 2 hours; then, the sequencing reactions are performed, which takes at least another hour, and the samples are run on the gel for several hours, the duration depending on the length of the fragment to be sequenced. By contrast, the present invention provides direct analysis of the PCR amplified material after brief transcription and fragmentation steps, saving days of time and labor.

An interesting clinical application for the characterization of heterozygous mutations with DNA chips is as follows. Individuals with germline cancer mutations have a very high risk for secondary tumors after treatment by irradiation. About 10% of all cancer patients may have germline mutations for p53 or other tumor suppressor genes. Thus, before deciding on a treatment modality, a physician could use the method and DNA chips of the invention to test for a germline suppressor gene mutation.

DNA Chips for Rational Therapeutic Management

The present invention also provides DNA chips that can be used by physicians to determine optimum therapeutic protocols by early, rapid detection of biologically mediated resistance to a therapeutic agent in a variety of disease states. The benefits of such DNA chips are many, as the chips will help physicians recognize health care cost savings, achieve rapid therapeutic benefits, limit administration of ineffective (due to the resistance) yet toxic drugs, monitor changes in pathogen resistance, and decrease pathogen acquisition of resistance. Important applications include the treatment of HIV, other infectious diseases, and cancer.

HIV has infected a large and expanding number of people, resulting in massive health care expenditures. HIV can rapidly become resistant to drugs used to treat the infection, primarily due to the action of the heterodimeric protein (51 kD and 66 kD) HIV reverse transcriptase (RT) encoded by the 1.7 kb pol gene. The high error rate (5–10 per round) of the RT protein is believed to account for the hypermutability of HIV. The nucleoside analogues, i.e., AZT, ddI, ddC, and d4T, commonly used to treat HIV infection are converted to nucleotide analogues by sequential phosphorylation in the cytoplasm of infected cells, where incorporation of the analogue into the viral DNA results in termination of viral replication, because the 5'→3' phosphodiester linkage cannot be completed. However, within after 6 months to 1 year of treatment, HIV typically mutates the RT gene so as to become incapable of incorporating the analogue and so resistant to treatment. Several known mutations are shown in tabular form below.

RT MUTATIONS ASSOCIATED WITH DRUG RESISTANCE

ANTI-VIRAL	CODON	aa CHANGE	nt CHANGE
AZT	67	Asp -> Asn	GAC -> AAC
AZT	70	Lys -> Arg	AAA -> AGA
AZT	215	Thr -> Phe or Tyr	ACC -> TTC or TAC
AZT	219	Lys -> Gln or Glu	AAA -> CAA or GAA
AZT	41	Met -> Leu	ATG -> TTG or CTG
ddI and ddC	184	Met -> Val	ATG -> GTG
ddI and ddC	74	Leu -> Val	
TIBO 82150	100	Leu -> Ile	

N.B. other mutations confer resistance to other drugs in vitro

The present invention provides DNA chips for detecting the multiple mutations in the HIV RT gene associated with resistance to different therapeutics. These DNA chips will enable physicians to monitor mutations over time and to change therapeutics if resistance develops. The DNA chip will provide redundant confirmation of conserved HIV RT and other gene sequences, and the probes on the chip will tile through, with overlap, in important mutational hot spot regions. The chip will optionally have probes that span the entire coding region of the RT and optionally the genes for other HIV proteins, such as coat proteins. HIV target nucleic acid can be isolated from blood samples (peripheral blood lymphocytes or PBMC) and amplified by PCR, primers for which are shown in the table below.

AMPLIFICATION OF TARGET

TARGET SIZE	PRIMER 1	PRIMER 2
1, 742bp	GTAGAATTCTGTTGACTCAGATTGG (SEQ ID. NO:292)	GATAAGCTTGGGCCTTATCTATTCCAT (SEQ ID. NO:294)
535bp	AAATCCATACAATACTCCAGTATTTGC (SEQ ID. NO:293)	ACCCATCCAAAGGAATGGAGGTTCTTTC (SEQ ID. NO:295)
323bp	Genbank#K02013 1889–1908	bases 2211–2192

The HIV RT gene chips of the invention, as well as the CF, mtDNA, and p53 DNA chips of the invention, illustrate the diverse application of the methods and probe arrays of the invention. The examples that follow describe methods for preparing nucleic acid targets from samples for application to the DNA chips of the invention and provide additional details of the methods of the invention.

EXAMPLES

I. VLSIPS™ Technology

As noted above, the VLSIPS™ technology is described in a number of patent publications and is preferred for making the oligonucleotide arrays of the invention. For completeness, a brief description of how this technology can be used to make and screen DNA chips is provided in this Example and the accompanying Figures. In the VLSIPS method, light is shone through a mask to activate functional (for oligonucleotides, typically an —OH) groups protected with a photoremovable protecting group on a surface of a solid support. After light activation, a nucleoside building block, itself protected with a photoremovable protecting group (at the 5'—OH), is coupled to the activated areas of the support. The process can be repeated, using different masks or mask orientations and building blocks, to prepare very dense arrays of many different oligonucleotide probes. The process is illustrated in FIG. 28; FIG. 29 illustrates how the process can be used to prepare "nucleoside combinatorials" or oligonucleotides synthesized by coupling all four nucleosides to form dimers, trimers, etc.

New methods for the combinatorial chemical synthesis of peptide, polycarbamate, and oligonucleotide arrays have recently been reported (see Fodor et al., 1991, *Science* 251: 767–773; Cho et al., 1993, *Science* 261: 1303–1305; and Southern et al., 1992, *Genomics* 13: 1008–10017, each of which is incorporated herein by reference). These arrays, or biological chips (see Fodor et al., 1993, *Nature* 364: 555–556, incorporated herein by reference), harbor specific chemical compounds at precise locations in a high-density, information rich format, and are a powerful tool for the study of biological recognition processes. A particularly exciting application of the array technology is in the field of DNA sequence analysis. The hybridization pattern of a DNA target to an array of shorter oligonucleotide probes is used to gain primary structure information of the DNA target. This format has important applications in sequencing by hybridization, DNA diagnostics and in elucidating the thermodynamic parameters affecting nucleic acid recognition.

Conventional DNA sequencing technology is a laborious procedure requiring electrophoretic size separation of labeled DNA fragments. An alternative approach, termed Sequencing By Hybridization (SBH), has been proposed (Lysov et al., 1988, *Dokl. Akad. Nauk SSSR* 303: 1508–1511; Bains et al., 1988, *J. Theor. Biol.* 135: 303–307; and Drmanac et al., 1989, *Genomics* 4: 114–128, incorporated herein by reference). This method uses a set of short oligonucleotide probes of defined sequence to search for complementary sequences on a longer target strand of DNA. The hybridization pattern is used to reconstruct the target DNA sequence. It is envisioned that hybridization analysis of large numbers of probes can be used to sequence long stretches of DNA. In immediate applications of this hybridization methodology, a small number of probes can be used to interrogate local DNA sequence.

The strategy of SBH can be illustrated by the following example. A 12-mer target DNA sequence, AGCCTAGCTGAA, (SEQ. ID NO:296) is mixed with a complete set of octanucleotide probes. If only perfect complementarity is considered, five of the 65,536 octamer probes -TCGGATCG, CGGATCGA, GGATCGAC, GATCGACT, and ATCGACTT will hybridize to the target. Alignment of the overlapping sequences from the hybridizing probes reconstructs the complement of the original 12-mer target:

```
TCGGATCG
 CGGATCGA
  GGATCGAC
   GATCGACT
    ATCGACTT
TCGGATCGACTT (SEQ. ID NO:297)
```

Hybridization methodology can be carried out by attaching target DNA to a surface. The target is interrogated with a set of oligonucleotide probes, one at a time (see Strezoska et al., 1991, *Proc. Natl. Acad. Sci. USA* 88: 10089–10093, and Drmanac et al., 1993, *Science* 260: 1649–1652, each of which is incorporated herein by reference). This approach can be implemented with well established methods of immobilization and hybridization detection, but involves a large number of manipulations. For example, to probe a sequence utilizing a full set of octanucleotides, tens of thousands, of hybridization reactions must be performed. Alternatively, SBH can be carried out by attaching probes to a surface in an array format where the identity of the probes at each site is known. The target DNA is then added to the array of probes. The hybridization pattern determined in a single experiment directly reveals the identity of all complementary probes.

As noted above, a preferred method of oligonucleotide probe array synthesis involves the use of light to direct the synthesis of oligonucleotide probes in high-density, miniaturized arrays. Photolabile 5'-protected N-acyl-deoxynucleoside phosphoramidites, surface linker chemistry, and versatile combinatorial synthesis strategies have been developed for this technology. Matrices of spatially-defined oligonucleotide probes have been generated, and the ability to use these arrays to identify complementary sequences has been demonstrated by hybridizing fluorescent labeled oligonucleotides to the DNA chips produced by the methods. The hybridization pattern demonstrates a high degree of base specificity and reveals the sequence of oligonucleotide targets.

The basic strategy for light-directed oligonucleotide synthesis (1) is outlined in FIG. 28. The surface of a solid support modified with photolabile protecting groups (X) is illuminated through a photolithographic mask, yielding reactive hydroxyl groups in the illuminated regions. A 3'-O-phosphoramidite activated deoxynucleoside (protected at the 5'-hydroxyl with a photolabile group) is then presented to the surface and coupling occurs at sites that were exposed to light. Following capping, and oxidation, the substrate is rinsed and the surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of products is obtained.

Light directed chemical synthesis lends itself to highly efficient synthesis strategies which will generate a maximum number of compounds in a minimum number of chemical steps. For example, the complete set of 4n polynucleotides (length n), or any subset of this set can be produced in only 4xn chemical steps. See FIG. 29. The patterns of illumination and the order of chemical reactants ultimately define the products and their locations. Because photolithography is used, the process can be miniaturized to generate high-density arrays of oligonucleotide probes. For an example of the nomenclature useful for describing such arrays, an array containing all possible octanucleotides of dA and dT is written as $(A+T)^8$. Expansion of this polynomial reveals the identity of all 256 octanucleotide probes from AAAAAAAA to TTTTTTTT. A DNA array composed of complete sets of dinucleotides is referred to as having a complexity of 2. The array given by (A+T+C+G)8 is the full 65,536 octanucleotide array of complexity four.

To carry out hybridization of DNA targets to the probe arrays, the arrays are mounted in a thermostatically controlled hybridization chamber. Fluorescein labeled DNA targets are injected into the chamber and hybridization is allowed to proceed for ½ to 2 hours. The surface of the matrix is scanned in an epifluorescence microscope (Zeiss Axioscop 20) equipped with photon counting electronics using 50–100 μW of 488 nm excitation from an Argon ion laser (Spectra Physics model 2020). All measurements are acquired with the target solution in contact with the probe matrix. Photon counts are stored and image files are presented after conversion to an eight bit image format. See FIG. 33.

When hybridizing a DNA target to an oligonucleotide array, N=Lt–(Lp–1) complementary hybrids are expected, where N is the number of hybrids, Lt is the length of the DNA target, and Lp is the length of the oligonucleotide probes on the array. For example, for an 11-mer hybridized to an octanucleotide array, N=4. Hybridizations with mismatches at positions that are 2 to 3 residues from either end of the probes will generate detectable signals. Modifying the above expression for N, one arrives at a relationship estimating the number of detectable hybridizations (Nd) for a DNA target of length Lt and an array of complexity C. Assuming an average of 5 positions giving signals above background: Nd=(1+5(C–1))[Lt–(Lp–1)].

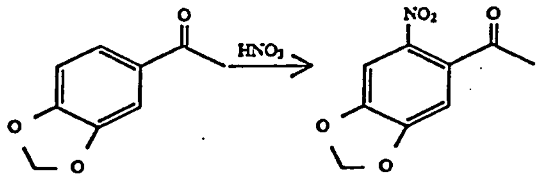
Arrays of oligonucleotides can be efficiently generated by light-directed synthesis and can be used to determine the identity of DNA target sequences. Because combinatorial strategies are used, the number of compounds increases exponentially while the number of chemical coupling cycles increases only linearly. For example, expanding the synthesis to the complete set of 4^8 (65,536) octanucleotides will add only four hours to the synthesis for the 16 additional cycles. Furthermore, combinatorial synthesis strategies can be implemented to generate arrays of any desired composition. For example, because the entire set of dodecamers (4^{12}) can be produced in 48 photolysis and coupling cycles (b^n compounds requires bxn cycles), any subset of the dodecamers (including any subset of shorter oligonucleotides) can be constructed with the correct lithographic mask design in 48 or fewer chemical coupling steps. In addition, the number of compounds in an array is limited only by the density of synthesis sites and the overall array size. Recent experiments have demonstrated hybridization to probes synthesized in 25 μm sites. At this resolution, the entire set of 65,536 octanucleotides can be placed in an array measuring 0.64 cm square, and the set of 1,048,576 dodecanucleotides requires only a 2.56 cm array.

Genome sequencing projects will ultimately be limited by DNA sequencing technologies. Current sequencing methodologies are highly reliant on complex procedures and require substantial manual effort. Sequencing by hybridization has the potential for transforming many of the manual efforts into more efficient and automated formats. Light-directed synthesis is an efficient means for large scale production of miniaturized arrays for SBH. The oligonucleotide arrays are not limited to primary sequencing applications. Because single base changes cause multiple changes in the hybridization pattern, the oligonucleotide arrays provide a powerful means to check the accuracy of previously elucidated DNA sequence, or to scan for changes within a sequence. In the case of octanucleotides, a single base change in the target DNA results in the loss of eight complements, and generates eight new complements. Matching of hybridization patterns may be useful in resolving sequencing ambiguities from standard gel techniques, or for rapidly detecting DNA mutational events. The potentially very high information content of light-directed oligonucleotide arrays will change genetic diagnostic testing. Sequence comparisons of hundreds to thousands of different genes will be assayed simultaneously instead of the current one, or few at a time format. Custom arrays can also be constructed to contain genetic markers for the rapid identification of a wide variety of pathogenic organisms.

Oligonucleotide arrays can also be applied to study the sequence specificity of RNA or protein-DNA interactions. Experiments can be designed to elucidate specificity rules of non Watson-Crick oligonucleotide structures or to investigate the use of novel synthetic nucleoside analogs for antisense or triple helix applications. Suitably protected RNA monomers may be employed for RNA synthesis. The oligonucleotide arrays should find broad application deducing the thermodynamic and kinetic rules governing formation and stability of oligonucleotide complexes.

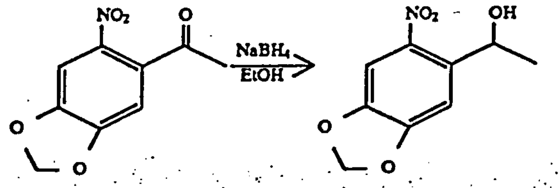
Other than the use of photoremovable protecting groups, the nucleoside coupling chemistry is very similar to that used routinely today for oligonucleotide synthesis. FIG. 30 shows the deprotection, coupling, and oxidation steps of a solid phase DNA synthesis method. FIG. 31 shows an illustrative synthesis route for the nucleoside building blocks used in the method. FIG. 32 shows a preferred photoremovable protecting group, MeNPOC, and how to prepare the group in active form. The procedures described below show how to prepare these reagents. The nucleoside building blocks are 5'-MeNPOC-THYMIDINE-3'-OCEP; 5'-MeNPOC-N^4-t-BUTYL PHENOXYACETYL-DEOXYCYTIDINE-3'-OCEP; 5'-MeNPOC-N^4-t-BUTYL PHENOXYACETYL-DEOXYGUANOSINE-3'-OCEP; and 5'-MeNPOC-N^4-t-BUTYL PHENOXYACETYL-DEOXYADENOSINE-3'-OCEP.

A. Preparation of 4, 5-methylenedioxy-2-nitroacetophenone



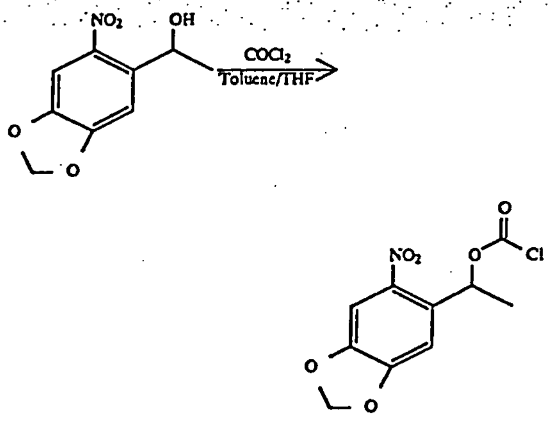
A solution of 50 g (0.305 mole) 3,4-methylenedioxyacetophenone (Aldrich) in 200 mL glacial acetic acid was added dropwise over 30 minutes to 700 mL of cold (2-4° C.) 70% HNO_3 with stirring (NOTE: the reaction will overheat without external cooling from an ice bath, which can be dangerous and lead to side products). At temperatures below 0° C., however, the reaction can be sluggish. A temperature of 3°–5° C. seems to be optimal). The mixture was left stirring for another 60 minutes at 3°–5° C., and then allowed to approach ambient temperature. Analysis by TLC (25% EtOAc in hexane) indicated complete conversion of the starting material within 1–2 hr. When the reaction was complete, the mixture was poured into ~3 liters of crushed ice, and the resulting yellow solid was filtered off, washed with water and then suction-dried. Yield ~53 g (84%), used without further purification.

B. Preparation of 1-(4,5-Methylenedioxy-2-nitrophenyl) ethanol



Sodium borohydride (10 g; 0.27 mol) was added slowly to a cold, stirring suspension of 53 g (0.25 mol) of 4,5-methylenedioxy-2-nitroacetophenone in 400 mL methanol. The temperature was kept below 10° C. by slow addition of the $NaBH_4$ and external cooling with an ice bath. Stirring was continued at ambient temperature for another two hours, at which time TLC (CH_2Cl_2) indicated complete conversion of the ketone. The mixture was poured into one liter of ice-water and the resulting suspension was neutralized with ammonium chloride and then extracted three times with 400 mL CH_2Cl_2 or EtOAc (the product can be collected by filtration and washed at this point, but it is somewhat soluble in water and this results in a yield of only ~60%). The combined organic extracts were washed with brine, then dried with $MgSO_4$ and evaporated. The crude product was purified from the main byproduct by dissolving it in a minimum volume of CH_2Cl_2 or THF(~175 ml) and then precipitating it by slowly adding hexane (1000 ml) while stirring (yield 51 g; 80% overall). It can also be recrystallized (eg., toluene-hexane), but this reduces the yield.

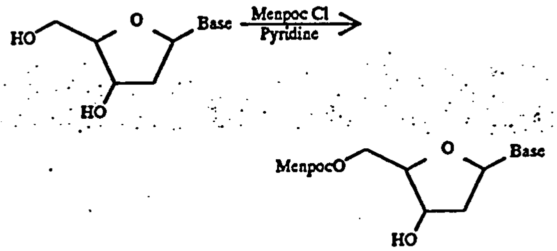
C. Preparation of 1-(4,5- methylenedioxy-2-nitrophenyl) ethyl chloroformate (MeNPOC-Cl)



Phosgene (500 mL of 20% w/v in toluene from Fluka: 965 mmole; 4 eq.) was added slowly to a cold, stirring solution of 50 g (237 mmole; 1 eq.) of 1-(4,5-methylenedioxy-2-nitrophenyl)ethanol in 400 mL dry THF. The solution was stirred overnight at ambient temperature at which point TLC (20% Et_2O/hexane) indicated >95% conversion. The mixture was evaporated (an oil-less pump with downstream aqueous NaOH trap is recommended to remove the excess phosgene) to afford a viscous brown oil. Purification was effected by flash chromatography on a short (9x13 cm) column of silica gel eluted with 20% Et_2O/hexane. Typically 55 g (85%) of the solid yellow MeNPOC-Cl is obtained by this procedure. The crude material has also been recrystallized in 2–3 crops from 1:1 ether/hexane. On this scale, ~100 ml is used for the first crop, with a few percent THF added to aid dissolution, and then cooling overnight at -20° C. (this procedure has not been optimized). The product should be stored dessicated at -20° C.

D. Synthesis of 5'-MeNPOC-2'-DEOXYNUCLEOSIDE-3'-(N,N-DIISOPROPYL 2-CYANOETHYL PHOSPHORAMIDITES

(1) 5'-MeNPOC-Nucleosides

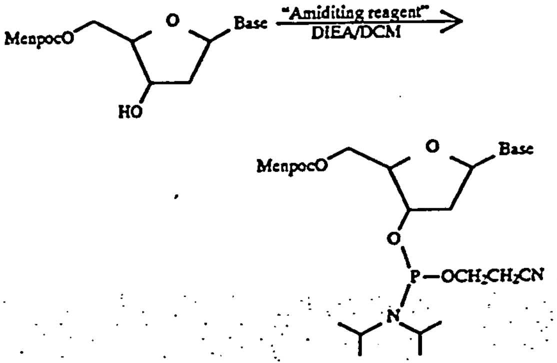


Base=THYMIDINE (T); N-4-ISOBUTYRYL 2'-DEOXYCYTIDINE (ibu-dC); N-2-PHENOXYACETYL 2'DEOXYGUANOSINE (PAC-dG); and N-6-PHENOXYACETYL 2'DEOXYADENOSINE (PAC-dA)

All four of the 5'-MeNPOC nucleosides were prepared from the base-protected 2'-deoxynucleosides by the following procedure. The protected 2'-deoxynucleoside (90 mmole) was dried by co-evaporating twice with 250 mL anhydrous pyridine. The nucleoside was then dissolved in

300 mL anhydrous pyridine (or 1:1 pyridine/DMF, for the dG^{PAC} nucleoside) under argon and cooled to −2° C. in an ice bath. A solution of 24.6 g (90 mmole) MeNPOC-Cl in 100 mL dry THP was then added with stirring over 30 minutes. The ice bath was removed, and the solution allowed to stir overnight at room temperature (TLC: 5–10% MeOH in CH_2Cl_2; two diastereomers). After evaporating the solvents under vacuum, the crude material was taken up in 250 mL ethyl acetate and extracted with saturated aqueous $NaHCO_3$ and brine. The organic phase was then dried over Na_2SO_4, filtered and evaporated to obtain a yellow foam. The crude products were finally purified by flash chromatography (9×30 cm silica gel column eluted with a stepped gradient of 2%–6% MeOH in CH_2Cl_2). Yields of the purified diastereomeric mixtures are in the range of 65–75%.

(2) 5'-MeNPOC-2'-DEOXYNUCLEOSIDE-3'-(N,N-DIISOPROPYL 2-CYANOETHYL PHOSPHORAMIDITES)



The four deoxynucleosides were phosphitylated using either 2-cyanoethyl-N,N-diisopropyl chlorophosphoramidite, or 2-cyanoethyl-N,N,N',N'-tetraisopropylphosphorodiamidite. The following is a typical procedure. Add 16.6 g (17.4 ml; 55 mmole) of 2-cyanoethyl-N,N,N',N'-tetraisopropylphosphorodiamidite to a solution of 50 mmole 5'-MeNPOC-nucleoside and 4.3 g (25 mmole) diisopropylammonium tetrazolide in 250 mL dry CH_2Cl_2 under argon at ambient temperature. Continue stirring for 4–16 hours (reaction monitored by TLC: 45:45:10 hexane/CH_2Cl_2/Et_3N). Wash the organic phase with saturated aqueous $NaHCO_3$ and brine, then dry over Na_2SO_4, and evaporate to dryness. Purify the crude amidite by flash chromatography (9×25 cm silica gel column eluted with hexane/CH_2Cl_2/TEA-45:45:10 for A, C, T; or 0:90:10 for G). The yield of purified amidite is about 90%.

II. PREPARATION OF LABELED DNA/HYBRIDIZATION TO ARRAY

1) PCR

PCR amplification reactions are typically conducted in a mixture composed of per reaction: 1 μl genomic DNA; 10 μl each primer (10 pmol/μl stocks); 10 μl 10×PCR buffer (100 mM Tris.Cl pH8.5, 500 mM KCl, 15 mM $MgCl_2$); 10 μl 2 mM dNTPs (made from 100 mM dNTP stocks); 2.5 U Taq polymerase (Perkin Elmer AmpliTaq™, 5 U/μl); and H_2O to 100 μl. The cycling conditions are usually 40 cycles (94° C. 45 sec, 55° C. 30 sec, 72° C. 60 sec) but may need to be varied considerably from sample type to sample type. These conditions are for 0.2 mL thin wall tubes in a Perkin Elmer 9600 thermocycler. See Perkin Elmer 1992/93 catalogue for 9600 cycle time information. Target, primer length and sequence composition, among other factors, may also affect parameters.

For products in the 200 to 1000 bp size range, check 2 μl of the reaction on a 1.5% 0.5×TBE agarose gel using an appropriate size standard (phiX174 cut with HaeIII is convenient). The PCR reaction should yield several picomoles of product. It is helpful to include a negative control (i.e., 1 μl TE instead of genomic DNA) to check for possible contamination. To avoid contamination, keep PCR products from previous experiments away from later reactions, using filter tips as appropriate. Using a set of working solutions and storing master solutions separately is helpful, so long as one does not contaminate the master stock solutions.

For simple amplifications of short fragments from genomic DNA it is, in general, unnecessary to optimize Mg^{2+} concentrations. A good procedure is the following: make a master mix minus enzyme; dispense the genomic DNA samples to individual tubes or reaction wells; add enzyme to the master mix; and mix and dispense the master solution to each well, using a new filter tip each time.

2) PURIFICATION

Removal of unincorporated nucleotides and primers from PCR samples can be accomplished using the Promega Magic PCR Preps DNA purification kit. One can purify the whole sample, following the instructions supplied with the kit (proceed from section IIIB, 'Sample preparation for direct purification from PCR reactions'). After elution of the PCR product in 50 μl of TE or H_2O, one centrifuges the eluate for 20 sec at 12,000 rpm in a microfuge and carefully transfers 45 μl to a new microfuge tube, avoiding any visible pellet. Resin is sometimes carried over during the elution step. This transfer prevents accidental contamination of the linear amplification reaction with 'Magic PCR' resin. Other methods, e.g. size exclusion chromatography, may also be used.

3) LINEAR AMPLIFICATION

In a 0.2 mL thin-wall PCR tube mix: 4 μl purified PCR product; 2 μl primer (10 pmol/μl); 4 μl 10×PCR buffer; 4 μl dNTPs (2 mM dA, dC, dG, 0.1 mM dT); 4 μl 0.1 mM dUTP; 1 μl 1 mM fluorescein dUTP (Amersham RPN 2121); 1 U Taq polymerase (Perkin Elmer, 5 U/μl); and add H_2O to 40 μl. Conduct 40 cycles (92° C. 30 sec, 55° C. 30 sec, 72° C. 90 sec) of PCR. These conditions have been used to amplify a 300 nucleotide mitochondrial DNA fragment but are generally applicable. Even in the absence of a visible product band on an agarose gel, there should still be enough product to give an easily detectable hybridization signal. If one is not treating the DNA with uracil DNA glycosylase (see Section 4), dUTP can be omitted from the reaction.

4) FRAGMENTATION

Purify the linear amplification product using the Promega Magic PCR Preps DNA purification kit, as per Section 2 above. In a 0.2 mL thin-wall PCR tube mix: 40 μl purified labeled DNA; 4 μl 10×PCR buffer; and 0.5 μl uracil DNA glycosylase (BRL 1U/μl). Incubate the mixture 15 min at 37° C., then 10 min at 97° C.; store at −20° C. until ready to use.

5) HYBRIDIZATION SCANNING & STRIPPING

A blank scan of the slide in hybridization buffer only is helpful to check that the slide is ready for use. The buffer is removed from the flow cell and replaced with 1 mL of (fragmented) DNA in hybridization buffer and mixed well. The scan is performed in the presence of the labeled target. FIG. 33 illustrates an illustrative detection system for scanning a DNA chip. A series of scans at 30 min intervals using a hybridization temperature of 25° C. yields a very clear signal, usually in at least 30 min to two hours, but it may be desirable to hybridize longer, i.e., overnight. Using a laser power of 50 μW and 50 μm pixels, one should obtain maximum counts in the range of hundreds to low thousands/

pixel for a new slide. When finished, the slide can be stripped using 50% formamide. rinsing well in deionized H_2O, blowing dry, and storing at room temperature.

III. PREPARATION OF LABELED RNA /HYBRIDIZATION TO ARRAY

1) TAGGED PRIMERS

The primers used to amplify the target nucleic acid should have promoter sequences if one desires to produce RNA from the amplified nucleic acid. Suitable promoter sequences are shown below and include:

(1) the T3 promoter sequence:

5'-CGGAATTAACCCTCACTAAAGG (SEQ. ID NO:298)

5'-AATTAACCCTCACTAAAGGGAG; (SEQ. ID NO:299)

(2) the T7 promoter sequence:

5' TAATACGACTCACTATAGGGAG; (SEQ. ID NO:300)

and (3) the SP6 promoter sequence:

5' ATTTAGGTGACACTATAGAA. (SEQ. ID NO:301)

The desired promoter sequence is added to the 5' end of the PCR primer. It is convenient to add a different promoter to each primer of a PCR primer pair so that either strand may be transcribed from a single PCR product.

Synthesize PCR primers so as to leave the DMT group on. DMT-on purification is unnecessary for PCR but appears to be important for transcription. Add 25 μl 0.5M NaOH to collection vial prior to collection of oligonucleotide to keep the DMT group on. Deprotect using standard chemistry—55° C. overnight is convenient.

HPLC purification is accomplished by drying down the oligonucleotides, resuspending in 1 mL 0.1M TEAA (dilute 2.0M stock in deionized water, filter through 0.2 micron filter) and filter through 0.2 micron filter. Load 0.5 mL on reverse phase HPLC (column can be a Hamilton PRP-1 semi-prep, #79426). The gradient is 0→50% CH_3CN over 25 min (program 0.2 μmol.prep.0–50, 25 min). Pool the desired fractions, dry down, resuspend in 200 μl 80% HAc. 30 min RT. Add 200 μl EtOH; dry down. Resuspend in 200 μl H_2O, plus 20 μl NaAc pH5.5, 600 μl EtOH. Leave 10 min on ice; centrifuge 12,000 rpm for 10 min in microfuge. Pour off supernatant. Rinse pellet with 1 mL EtOH, dry, resuspend in 200 μl H2O. Dry, resuspend in 200 μl TE. Measure A260, prepare a 10 pmol/μl solution in TE (10 mM Tris.Cl pH 8.0, 0.1 mM EDTA). Following HPLC purification of a 42 mer, a yield in the vicinity of 15 nmol from a 0.2 μmol scale synthesis is typical.

2) GENOMIC DNA PREPARATION

For obtaining genomic DNA from human hair, one can extract as few as 5 hairs, including hair roots. On a clean and sterile surface, one places the hair on a piece of parafilm, and after wiping a new razor blade with EtOH cutting off the roots, the roots are transferred to a 1.5 mL microfuge tube using a pair of Millipore forceps cleaned with EtOH. Add 500 μl (10 mM Tris.Cl pH8.0, 10 mM EDTA, 100 mM NaCl, 2% (w/v) SDS, 40 mM DTT, filter sterilized) to the sample. Add 1.25 μl 20 mg/ml proteinase K (Boehringer) Incubate at 55° C. for 2 hours, vortexing once or twice. Perform 2x0.5 mL 1:1 phenol:$CHCl_3$ extractions. After each extraction, centrifuge 12,000 rpm 5 min in a microfuge and recover 0.4 mL supernatant. Add 35 μl NaAc pH5.2 plus 1 mL EtOH. Place sample on ice 45 min; then centrifuge 12,000 rpm 30 min, rinse, air dry 30 min, and resuspend in 100 μl TE.

3) PCR

PCR is performed in a mixture containing, per reaction: 1 μl genomic DNA; 4 μl each primer (10 pmol/μl stocks); 4 μl 10 ×PCR buffer (100 mM Tris.Cl pH8.5, 500 mM KCl, 15 mM $MgCl_2$); 4 μl 2 mM dNTPs (made from 100 mM dNTP stocks); 1 U Taq polymerase (Perkin Elmer, 5 U/μl); H_2O to 40 μl. About 40 cycles (94° C. 30 sec, 55° C. 30 sec, 72° C. 30 sec) are performed, but cycling conditions may need to be varied. These conditions are for 0.2 mL thin wall tubes in Perkin Elmer 9600. For products in the 200 to 1000 bp size range, check 2 μl of the reaction on a 1.5% 0.5×TBE agarose gel using an appropriate size standard. For larger or smaller volumes (20–100 μl), one can use the same amount of genomic DNA but adjust the other ingredients accordingly.

4) IN VITRO TRANSCRIPTION

Mix: 3 μl PCR product; 4 μl 5xbuffer; 2 μl DTT; 2.4 μl 10 mM rNTPs (100 mM solutions from Pharmacia); 0.48 μl 10 mM fluorescein-UTP (Fluorescein-12-UTP, 10 mM solution, from Boehringer Mannheim); 0.5 μl RNA polymerase (Promega T3 or T7 RNA polymerase); and add H_2O to 20 μl. Incubate at 37° C. for 3 h. Check 2 μl of the reaction on a 1.5% 0.5×TBE agarose gel using a size standard. 5xbuffer is 200 mM Tris pH 7.5, 30 mM $MgCl_2$, 10 mM spermidine, 50 mM NaCl, and 100 mM DTT (supplied with enzyme). The PCR product needs no purification and can be added directly to the transcription mixture. A 20 μl reaction is suggested for an initial test experiment and hybridization; a 100 μl reaction is considered "preparative" scale (the reaction can be scaled up to obtain more target). The amount of PCR product to add is variable; typically a PCR reaction will yield several picomoles of DNA. If the PCR reaction does not produce that much target, then one should increase the amount of DNA added to the transcription reaction (as, well as optimize the PCR). The ratio of fluorescein-UTP to UTP suggested above is 1:5, but ratios from 1:3 to 1:10—all work well. One can also label with biotin-UTP and detect with streptavidin-FITC to obtain similar results as with fluorescein-UTP detection.

For nondenaturing agarose gel electrophoresis of RNA, note that the RNA band will normally migrate somewhat faster than the DNA template band, although sometimes the two bands will comigrate. The temperature of the gel can effect the migration of the RNA band. The RNA produced from in vitro transcription is quite stable and can be stored for months (at least) at −20° C. without any evidence of degradation. It can be stored in unsterilized 6xSSPE 0.1% triton X- 100 at −20° C. for days (at least) and reused twice (at least) for hybridization, without taking any special precautions in preparation or during use. RNase contamination should of course be avoided. When extracting RNA from cells, it is preferable to work very rapidly and to use strongly denaturing conditions. Avoid using glassware previously contaminated with RNases. Use of new disposable plasticware (not necessarily sterilized) is preferred, as new plastic tubes, tips, etc., are essentially RNase free. Treatment with DEPC or autoclaving is typically not unnecessary.

5) FRAGMENTATION

In a 0.2 mL thin-wall PCR tube mix: 18 μl RNA (direct from transcription reaction—no purification required); 18 μl H_2O; and 4 μl 1M Tris.Cl pH9.0. Incubate at 99.9° C. for 60 min. Add to 1 mL hybridization buffer and store at −20° C. until ready to use. The alkaline hydrolysis step is very reliable. The hydrolysed target can be stored at −20° C. in 6xSSPE/0.1% Triton X-100 for at least several days prior to use and can also be reused.

6) HYBRIDIZATION SCANNING, & STRIPPING

A blank scan of the slide in hybridization buffer only is helpful to check that the slide is ready for use. The buffer is removed from the flow cell and replaced with 1 mL of (hydrolysed) RNA in hybridization buffer and mixed well. Incubate for 15–30 min at 18° C. Remove the hybridization solution, which can be saved for subsequent experiments. Rinse the flow cell 4–5 times with fresh changes of 6xSSPE/0.1% Triton X-100, equilibrated to 18° C. The rinses can be

AK

US006156501A

United States Patent [19]

McGall et al.

[11] **Patent Number: 6,156,501**

[45] **Date of Patent: *Dec. 5, 2000**

[54] **ARRAYS OF MODIFIED NUCLEIC ACID PROBES AND METHODS OF USE**

[75] Inventors: Glenn Hugh McGall; Charles Garrett Miyada, both of Mountain View; Maureen T. Cronin, Los Altos; Jennifer Dee Tan, Newark; Mark S. Chee, Palo Alto, all of Calif.

[73] Assignee: Affymetrix, Inc., Santa Clara, Calif.

[*] Notice: This patent issued on a continued prosecution application filed under 37 CFR 1.53(d), and is subject to the twenty year patent term provisions of 35 U.S.C. 154(a)(2).

[21] Appl. No.: 08/630,427

[22] Filed: Apr. 3, 1996

Related U.S. Application Data

[63] Continuation-in-part of application No. 08/440,742, May 10, 1995, abandoned, which is a continuation-in-part of application No. PCT/US94/12305, Oct. 26, 1994, which is a continuation-in-part of application No. 08/284,064, Aug. 2, 1994, abandoned, which is a continuation-in-part of application No. 08/143,312, Oct. 26, 1993, abandoned.

[51] Int. Cl.⁷ C12Q 1/68; C07H 21/00

[52] U.S. Cl. 435/6; 422/50; 422/68.1; 435/283.1; 435/287.1; 435/287.2; 435/288.3; 435/288.7; 435/289.1; 435/299.1; 435/305.1; 436/501; 536/22.1; 536/24.1; 536/24.3; 536/24.31; 536/24.32; 536/24.33

[58] Field of Search 422/50, 68.1; 435/6, 435/810, 283.1, 287.1, 287.2, 288.3, 288.7, 289.1, 299.1, 305.1; 436/501; 536/23.1, 24.1, 22.1, 24.3–24.33; 935/77, 78

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,002,867	3/1991	Macevicz	435/6
5,143,854	9/1992	Pirrung et al.	436/518
5,200,051	4/1993	Cozzette et al.	204/403
5,202,231	4/1993	Drmanac et al.	435/6
5,217,866	6/1993	Summerton et al.	435/6
5,412,087	5/1995	McGall et al.	536/24.3
5,474,796	12/1995	Brennan	427/2.13
5,484,908	1/1996	Froehler et al.	536/24.31
5,527,681	6/1996	Holmes	435/6
5,556,752	9/1996	Lockhart et al.	435/6
5,556,961	9/1996	Foote et al.	536/27.1
5,604,097	2/1997	Brenner	435/6
5,800,992	9/1998	Fodor et al.	435/6
5,821,060	10/1998	Arlinghaus et al.	435/6
B1 4,683,202	11/1990	Mullis	435/91

FOREIGN PATENT DOCUMENTS

8605518	9/1986	WIPO .	
WO 89/10977	11/1989	WIPO .	
WO 89/11548	11/1989	WIPO .	
WO90/04652	5/1990	WIPO	C12Q 1/68
WO 92/10092	6/1992	WIPO .	
WO 93/17126	9/1993	WIPO .	

OTHER PUBLICATIONS

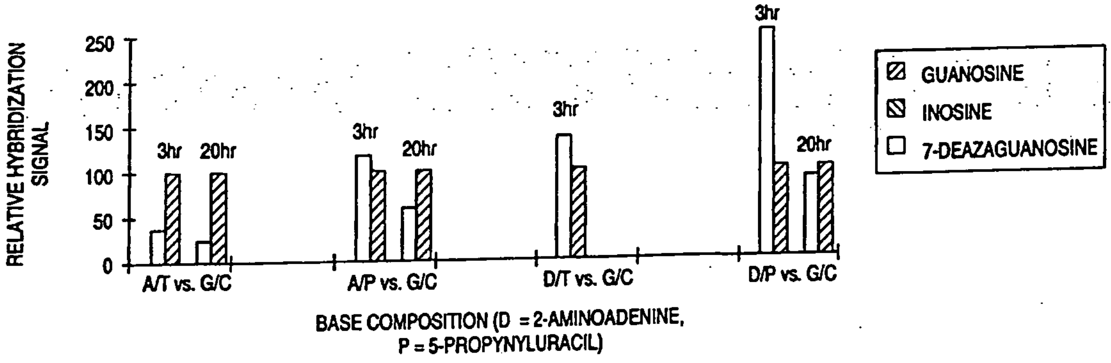
Durland, Ross H., et al. (1995) "Selective Binding of Pyrido[2, 3–d]pyrimidine 2'-Deoxyribonucleoside to AT Base Pairs in Antiparallel Triple Helices", *Bioconjugate Chem* 6:278–282.

Primary Examiner—Ardin H. Marschel
Attorney, Agent, or Firm—Townsend and Townsend and Crew

[57] **ABSTRACT**

Oligonucleotide analogue arrays attached to solid substrates and methods related to the use thereof are provided. The oligonucleotide analogues hybridize to nucleic acids with either higher or lower specificity than corresponding unmodified oligonucleotides. Target nucleic acids which comprise nucleotide analogues are bound to oligonucleotide and oligonucleotide analogue arrays.

72 Claims, 5 Drawing Sheets



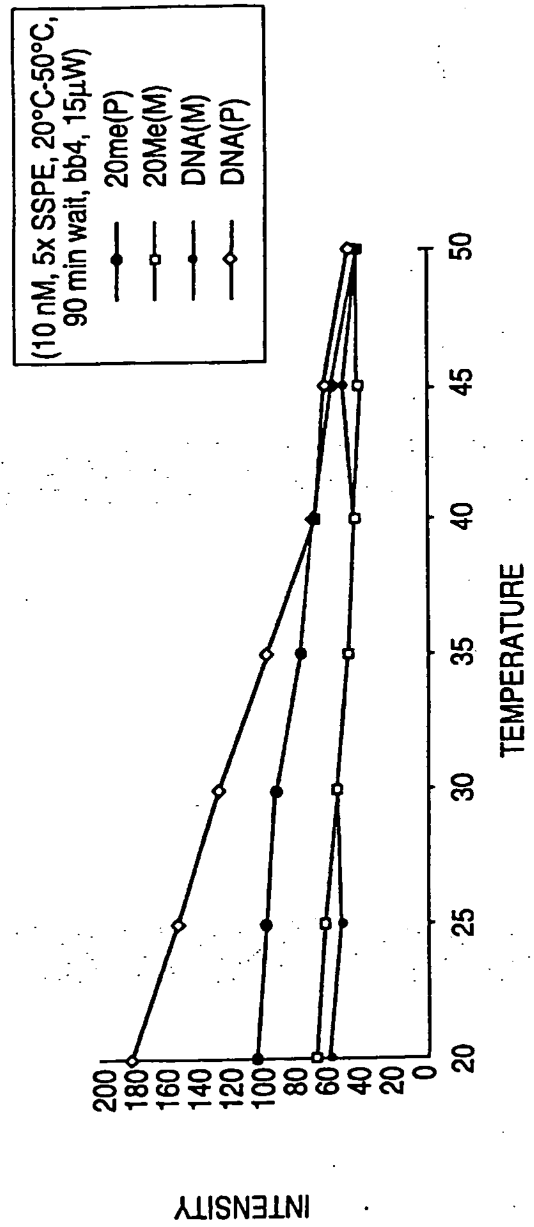

(10 nM, 5x SSPE, 20°C-50°C,
90 min wait, bb4, 15µW)
20me(P)
20Me(M)
DNA(M)
DNA(P)
INTENSITY
200
180
160
140
120
100
80
60
40
20
0
20
25
30
35
40
45
50
TEMPERATURE

FIG. 1A

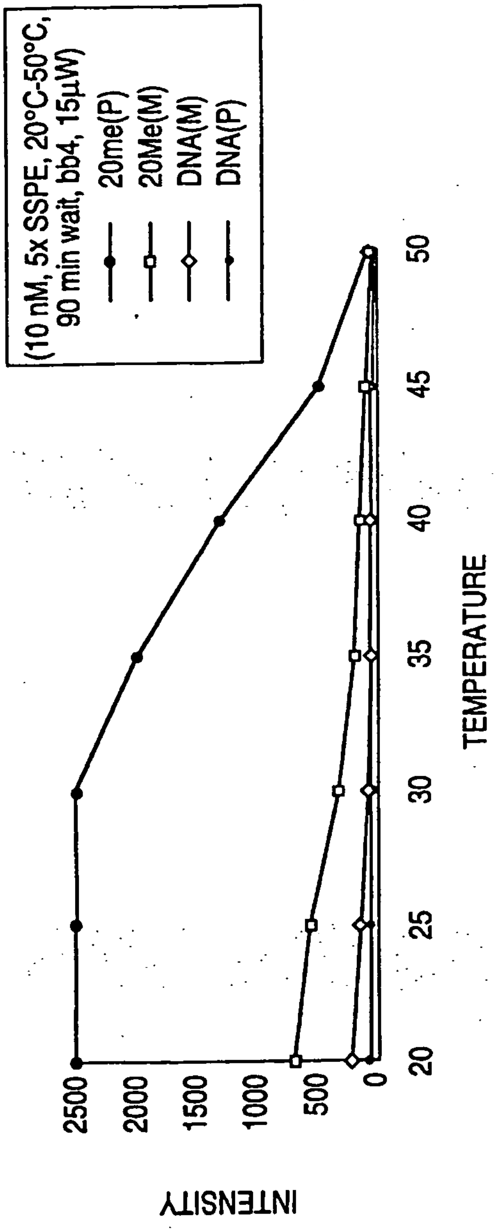

(10 nM, 5x SSPE, 20°C-50°C,
90 min wait, bb4, 15µW)
20me(P)
20Me(M)
DNA(M)
DNA(P)
INTENSITY
2500
2000
1500
1000
500
0
20
25
30
35
40
45
50
TEMPERATURE

FIG. 1B

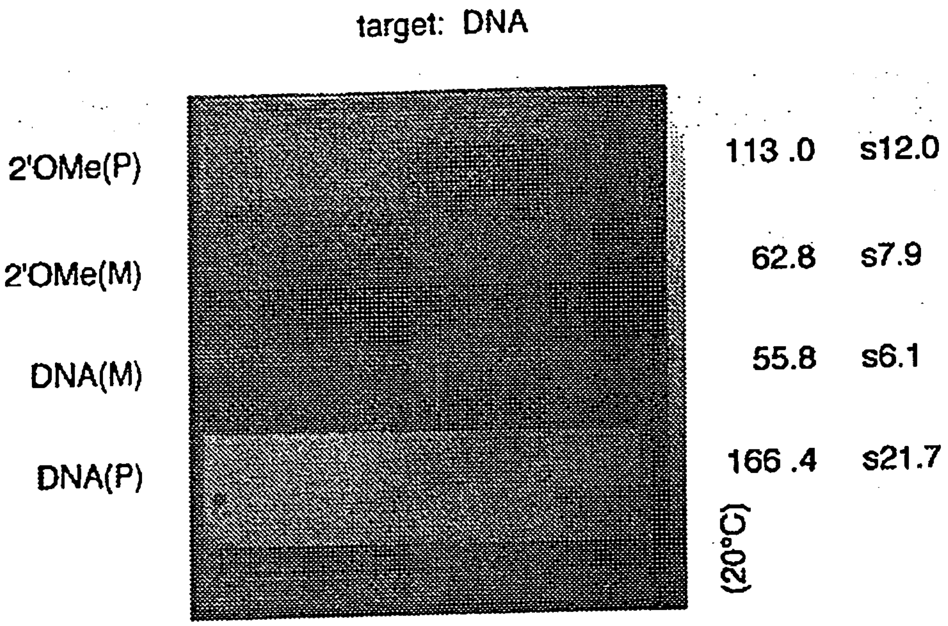

target: DNA
2'OMe(P)
2'OMe(M)
DNA(M)
DNA(P)
113 .0 s12.0
62.8 s7.9
55.8 s6.1
166 .4 s21.7
(20°C)

FIG. 1C

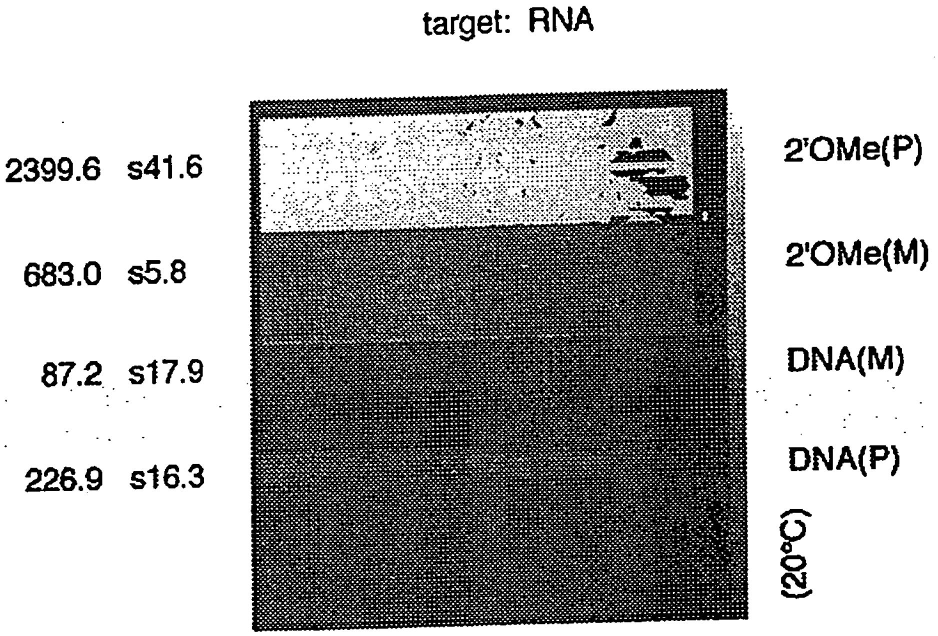

target: RNA
2399.6 s41.6
683.0 s5.8
87.2 s17.9
226.9 s16.3
2'OMe(P)
2'OMe(M)
DNA(M)
DNA(P)
(20°C)

FIG. 1D

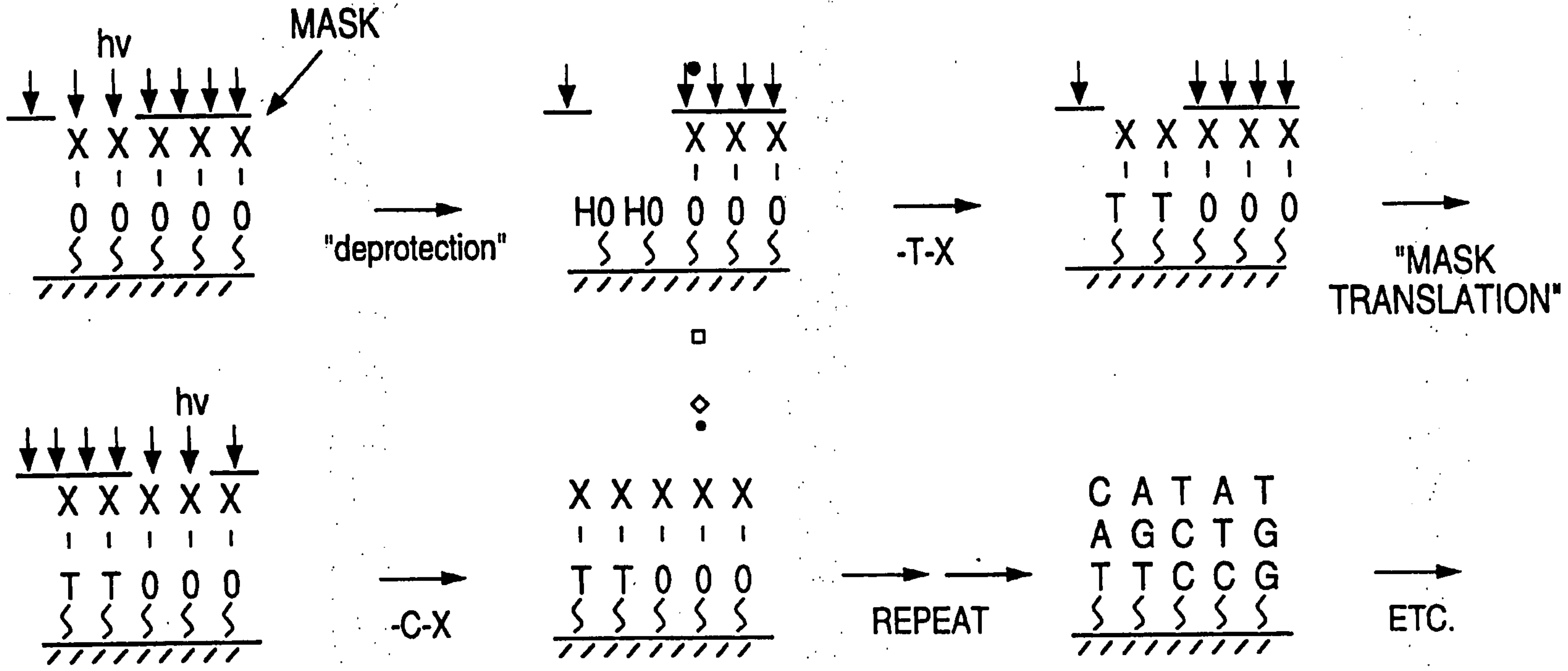

hv
MASK
X X X X X
0 0 0 0 0
"deprotection"
X X X
H0 H0 0 0 0
-T-X
X X X X X
T T 0 0 0
"MASK TRANSLATION"
hv
X X X X X
T T 0 0 0
-C-X
X X X X X
T T 0 0 0
REPEAT
C A T A T
A G C T G
T T C C G
ETC.

FIG. 2

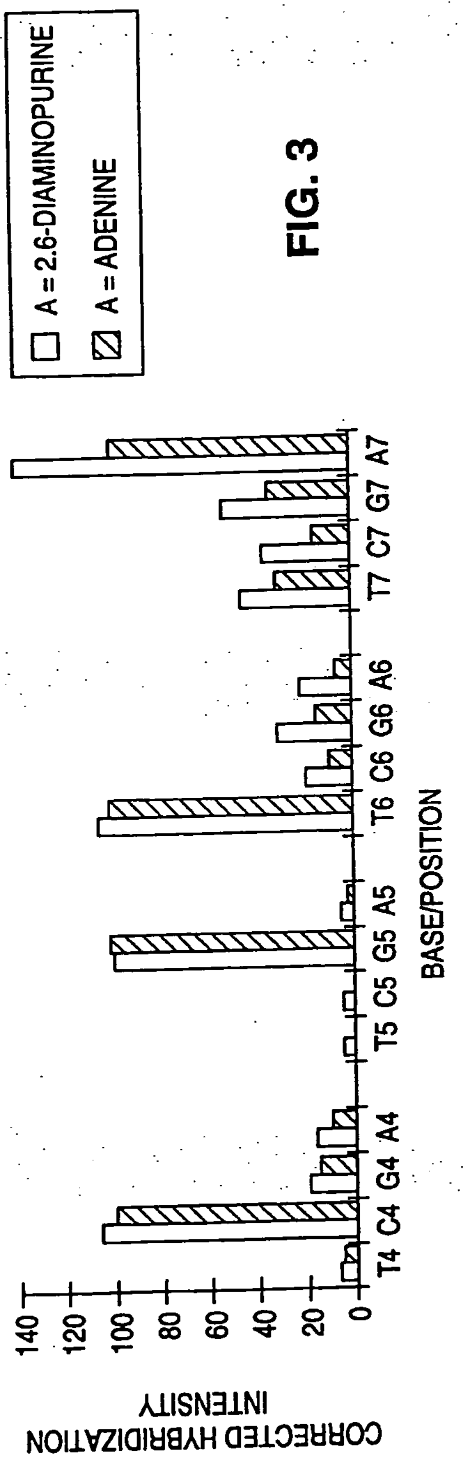

□ A = 2.6-DIAMINOPURINE
▨ A = ADENINE
CORRECTED HYBRIDIZATION INTENSITY
140
120
100
80
60
40
20
0
T4
C4
G4
A4
T5
C5
G5
A5
T6
C6
G6
A6
T7
C7
G7
A7
BASE/POSITION

FIG. 3

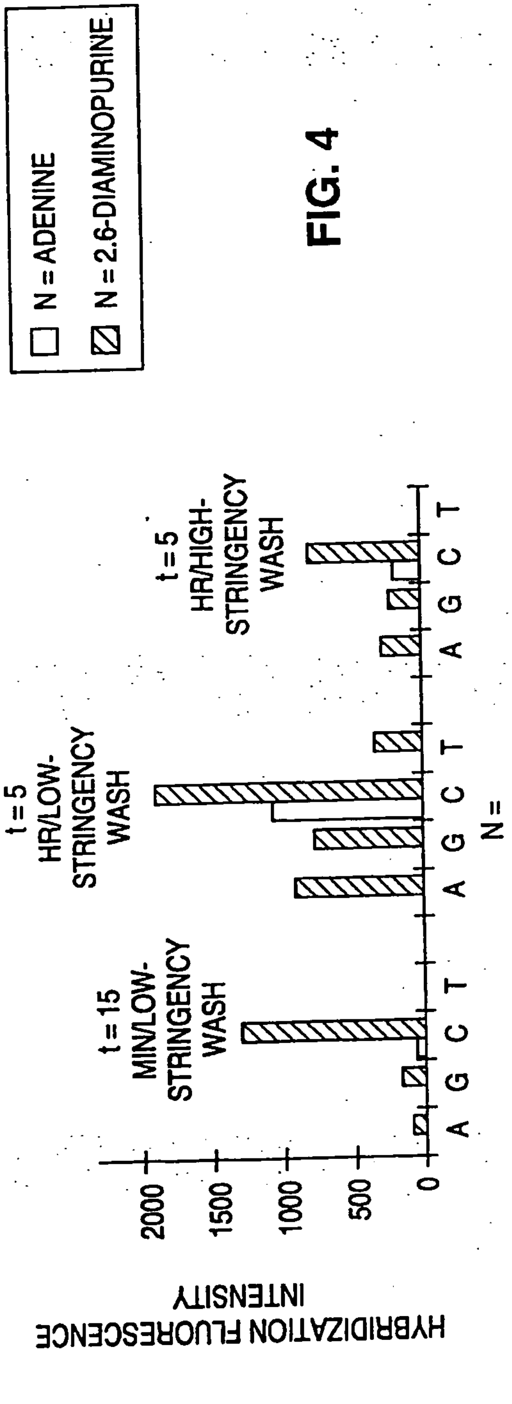

□ N = ADENINE
▨ N = 2.6-DIAMINOPURINE
HYBRIDIZATION FLUORESCENCE INTENSITY
2000
1500
1000
500
0
t = 15 MIN/LOW-STRINGENCY WASH
t = 5 HR/LOW-STRINGENCY WASH
t = 5 HR/HIGH-STRINGENCY WASH
A
G
C
T
A
G
C
T
A
G
C
T
N =

FIG. 4

FIG. 5

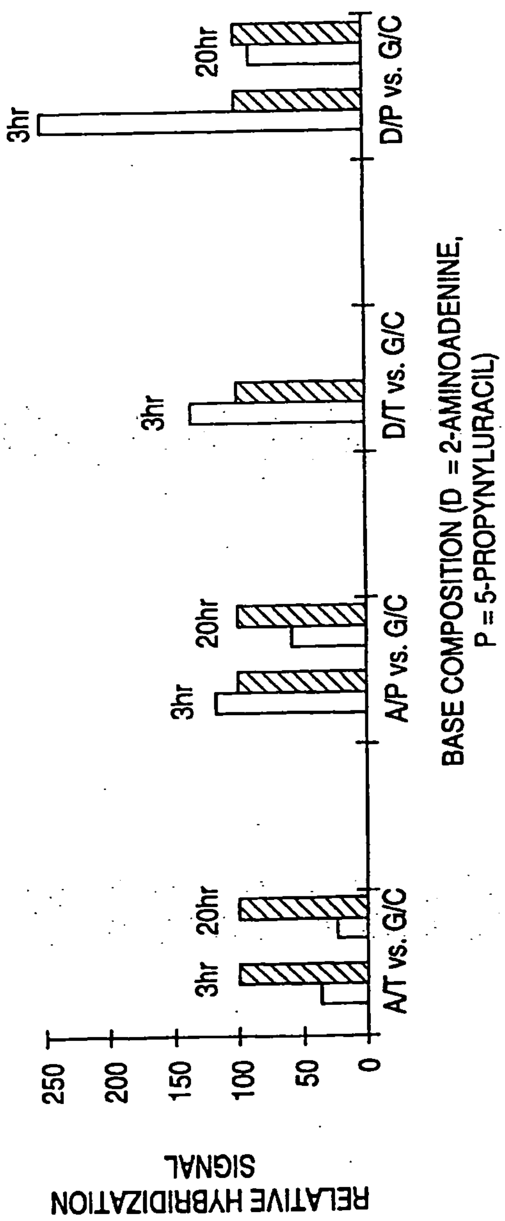
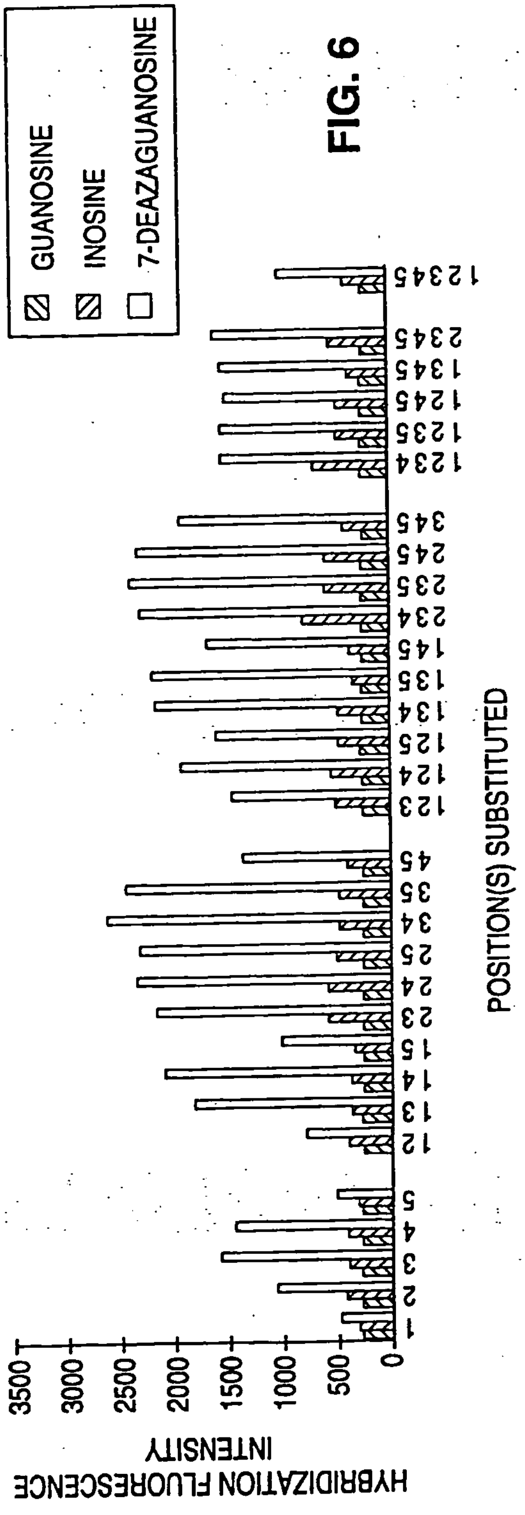

RELATIVE HYBRIDIZATION SIGNAL
250
200
150
100
50
0
3hr
20hr
A/T vs. G/C
A/P vs. G/C
D/T vs. G/C
D/P vs. G/C
BASE COMPOSITION (D = 2-AMINOADENINE, P = 5-PROPYNYLURACIL)

FIG. 6


▧ GUANOSINE
▨ INOSINE
□ 7-DEAZAGUANOSINE
HYBRIDIZATION FLUORESCENCE INTENSITY
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
1 2
1 3
1 4
1 5
2 3
2 4
2 5
3 4
3 5
4 5
1 2 3
1 2 4
1 2 5
1 3 4
1 3 5
1 4 5
2 3 4
2 3 5
2 4 5
3 4 5
1 2 3 4
1 2 3 5
1 2 4 5
1 3 4 5
2 3 4 5
1 2 3 4 5
POSITION(S) SUBSTITUTED

ARRAYS OF MODIFIED NUCLEIC ACID PROBES AND METHODS OF USE

CROSS-REFERENCE TO RELATED APPLICATION

This application is a continuation-in-part of U.S. Ser. No. 08/440,742 filed May 10, 1995 abandoned, which is a continuation-in-part of PCT application (designating the United States) SN PCT/US94/12305 filed Oct. 26, 1994, which is a continuation-in-part of U.S. Ser. No. 08/284,064 filed Aug. 2, 1994 abandoned, which is a continuation-in-part of U.S. Ser. No 08/143,312 filed Oct. 26, 1993 abandoned, each of which is incorporated herein by reference in its entirety for all purposes.

FIELD OF THE INVENTION

The present invention provides probes comprised of nucleotide analogues immobilized in arrays on solid substrates for analyzing molecular interactions of biological interest, and target nucleic acids comprised of nucleotide analogues. The invention therefore relates to the molecular interaction of polymers immobilized on solid substrates including related chemistry, biology, and medical diagnostic uses.

BACKGROUND OF THE INVENTION

The development of very large scale immobilized polymer synthesis (VLSIPS™) technology provides pioneering methods for arranging large numbers of oligonucleotide probes in very small arrays. See, U.S. application Ser. No. 07/805,727 now U.S. Pat. No. 5,424,186 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated herein by reference for all purposes. U.S. patent application Ser. No. 08/082,937, filed Jun. 25, 1993, and incorporated herein for all purposes, describes methods for making arrays of oligonucleotide probes that are used, e.g., to determine the complete sequence of a target nucleic acid and/or to detect the presence of a nucleic acid with a specified sequence.

VLSIPS™ technology provides an efficient means for large scale production of miniaturized oligonucleotide arrays for sequencing by hybridization (SBH), diagnostic testing for inherited or somatically acquired genetic diseases, and forensic analysis. Other applications include determination of sequence specificity of nucleic acids, protein-nucleic acid complexes and other polymer-polymer interactions.

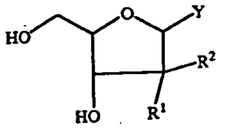
SUMMARY OF THE INVENTION

The present invention provides arrays of oligonucleotide analogues attached to solid substrates. Oligonucleotide analogues have different hybridization properties than oligonucleotides based upon naturally occurring nucleotides. By incorporating oligonucleotide analogues into the arrays of the invention, hybridization to a target nucleic acid is optimized.

The oligonucleotide analogue arrays have virtually any number of different members, determined largely by the number or variety of compounds to be screened against the array in a given application. In one group of embodiments, the array has from 10 up to 100 oligonucleotide analogue members. In other groups of embodiments, the arrays have between 100 and 10,000 members, and in yet other embodiments the arrays have between 10,000 and 1,000,0000 members. In preferred embodiments, the array will have a density of more than 100 members at known locations per cm^2, or more preferably, more than 1000 members per cm^2. In some embodiments, the arrays have a density of more than 10,000 members per cm^2.

The solid substrate upon which the array is constructed includes any material upon which oligonucleotide analogues are attached in a defined relationship to one another, such as beads, arrays, and slides. Especially preferred oligonucleotide analogues of the array are between about 5 and about 20 nucleotides, nucleotide analogues or a mixture thereof in length.

In one group of embodiments, nucleoside analogues incorporated into the oligonucleotide analogues of the array will have the chemical formula:



wherein R^1 and R^2 are independently selected from the group consisting of hydrogen, methyl, hydroxy, alkoxy (e.g., methoxy, ethoxy, propoxy, allyloxy, and propargyloxy), alkylthio, halogen (Fluorine, Chlorine, and Bromine), cyano, and azido, and wherein Y is a heterocyclic moiety, e.g., a base selected from the group consisting of purines, purine analogues, pyrimidines, pyrimidine analogues, universal bases (e.g., 5-nitroindole) or other groups or ring systems capable of forming one or more hydrogen bonds with corresponding moieties on alternate strands within a double- or triple-stranded nucleic acid or nucleic acid analogue, or other groups or ring systems capable of forming nearest-neighbor base-stacking interactions within a double- or triple-stranded complex. In other embodiments, the oligonucleotide analogues are not constructed from nucleosides, but are capable of binding to nucleic acids in solution due to structural similarities between the oligonucleotide analogue and a naturally occurring nucleic acid. An example of such an oligonucleotide analogue is a peptide nucleic acid or polyamide nucleic acid in which bases which hydrogen bond to a nucleic acid are attached to a polyamide backbone.

The present invention also provides target nucleic acids hybridized to oligonucleotide arrays. In the target nucleic acids of the invention, nucleotide analogues are incorporated into the target nucleic acid, altering the hybridization properties of the target nucleic acid to an array of oligonucleotide probes. Typically, the oligonucleotide probe arrays also comprise nucleotide analogues.

The target nucleic acids are typically synthesized by providing a nucleotide analogue as a reagent during the enzymatic copying of a nucleic acid. For instance, nucleotide analogues are incorporated into polynucleic acid analogues using taq polymerase in a PCR reaction. Thus, a nucleic acid containing a sequence to be analyzed is typically amplified in a PCR or RNA amplification procedure with nucleotide analogues, and the resulting target nucleic acid analogue amplicon is hybridized to a nucleic acid analogue array.

Oligonucleotide analogue arrays and target nucleic acids are optionally composed of oligonucleotide analogues which are resistant to hydrolysis or degradation by nuclease enzymes such as RNAase A. This has the advantage of providing the array or target nucleic acid with greater longevity by rendering it resistant to enzymatic degradation.

For example, analogues comprising 2'-O-methyloligoribonucleotides are resistant to RNAase A.

Oligonucleotide analogue arrays are optionally arranged into libraries for screening compounds for desired characteristics, such as the ability to bind a specified oligonucleotide analogue, or oligonucleotide analogue-containing structure. The libraries also include oligonucleotide analogue members which form conformationally-restricted probes, such as unimolecular double-stranded probes or unimolecular double-stranded probes which present a third chemical structure of interest. For instance, the array of oligonucleotide analogues optionally include a plurality of different members, each member having the formula: $Y—L^1—X^1—L^2—X^2$, wherein Y is a solid substrate, X^1 and X^2 are complementary oligonucleotides containing at least one nucleotide analogue, L^1 is a spacer, and L^2 is a linking group having sufficient length such that X^1 and X^2 form a double-stranded oligonucleotide. An array of such members comprise a library of unimolecular double-stranded oligonucleotide analogues. In another embodiment, the members of the array of oligonucleotide are arranged to present a moiety of interest within the oligonucleotide analogue probes of the array. For instance, the arrays are optionally conformationally restricted, having the formula $—X^{11}—Z—X^{12}$, wherein X^{11} and X^{12} are complementary oligonucleotides or oligonucleotide analogues and Z is a chemical structure comprising the binding site of interest.

Oligonucleotide analogue arrays are synthesized on a solid substrate by a variety of methods, including light-directed chemical coupling, and selectively flowing synthetic reagents over portions of the solid substrate. The solid substrate is prepared for synthesis or attachment of oligonucleotides by treatment with suitable reagents. For example, glass is prepared by treatment with silane reagents.

The present invention provides methods for determining whether a molecule of interest binds members of the oligonucleotide analogue array. For instance, in one embodiment, a target molecule is hybridized to the array and the resulting hybridization pattern is determined. The target molecule includes genomic DNA, cDNA, unspliced RNA, mRNA, and rRNA, nucleic acid analogues, proteins and chemical polymers. The target molecules are optionally amplified prior to being hybridized to the array, e.g., by PCR, LCR, or cloning methods.

The oligonucleotide analogue members of the array used in the above methods are synthesized by any described method for creating arrays. In one embodiment, the oligonucleotide analogue members are attached to the solid substrate, or synthesized on the solid substrate by light-directed very large scale immobilized polymer synthesis, e.g., using photo-removable protecting groups during synthesis. In another embodiment, the oligonucleotide members are attached to the solid substrate by forming a plurality of channels adjacent to the surface of said substrate, placing selected monomers in said channels to synthesize oligonucleotide analogues at predetermined portions of selected regions, wherein the portion of the selected regions comprise oligonucleotide analogues different from oligonucleotide analogues in at least one other of the selected regions, and repeating the steps with the channels formed along a second portion of the selected regions. The solid substrate is any suitable material as described above, including beads, slides, and arrays, each of which is constructed from, e.g., silica, polymers and glass.

DEFINITIONS

An "Oligonucleotide" is a nucleic acid sequence composed of two or more nucleotides. An oligonucleotide is optionally derived from natural sources, but is often synthesized chemically. It is of any size. An "oligonucleotide analogue" refers to a polymer with two or more monomeric subunits, wherein the subunits have some structural features in common with a naturally occurring oligonucleotide which allow it to hybridize with a naturally occurring oligonucleotide in solution. For instance, structural groups are optionally added to the ribose or base of a nucleoside for incorporation into an oligonucleotide, such as a methyl or allyl group at the 2'-O position on the ribose, or a fluoro group which substitutes for the 2'-O group, or a bromo group on the ribonucleoside base. The phosphodiester linkage, or "sugar-phosphate backbone" of the oligonucleotide analogue is substituted or modified, for instance with methyl phosphonates or O-methyl phosphates. Another example of an oligonucleotide analogue for purposes of this disclosure includes "peptide nucleic acids" in which native or modified nucleic acid bases are attached to a polyamide backbone. Oligonucleotide analogues optionally comprise a mixture of naturally occurring nucleotides and nucleotide analogues. However, an oligonucleotide which is made entirely of naturally occurring nucleotides (i.e., those comprising DNA or RNA), with the exception of a protecting group on the end of the oligonucleotide, such as a protecting group used during standard nucleic acid synthesis is not considered an oligonucleotide analogue for purposes of this invention.

A "nucleoside" is a pentose glycoside in which the aglycone is a heterocyclic base; upon the addition of a phosphate group the compound becomes a nucleotide. The major biological nucleosides are β-glycoside derivatives of D-ribose or D-2-deoxyribose. Nucleotides are phosphate esters of nucleosides which are generally acidic in solution due to the hydroxy groups on the phosphate. The nucleosides of DNA and RNA are connected together via phosphate units attached to the 3' position of one pentose and the 5' position of the next pentose. Nucleotide analogues and/or nucleoside analogues are molecules with structural similarities to the naturally occurring nucleotides or nucleosides as discussed above in the context of oligonucleotide analogues.

A "nucleic acid reagent" utilized in standard automated oligonucleotide synthesis typically caries a protected phosphate on the 3' hydroxyl of the ribose. Thus, nucleic acid reagents are referred to as nucleotides, nucleotide reagents, nucleoside reagents, nucleoside phosphates, nucleoside-3'-phosphates, nucleoside phosphoramidites, phosphoramidites, nucleoside phosphonates, phosphonates and the like. It is generally understood that nucleotide reagents carry a reactive, or activatible, phosphoryl or phosphonyl moiety in order to form a phosphodiester linkage.

A "protecting group" as used herein, refers to any of the groups which are designed to block one reactive site in a molecule while a chemical reaction is carried out at another reactive site. More particularly, the protecting groups used herein are optionally any of those groups described in Greene, et al., *Protective Groups In Organic Chemistry*, 2nd Ed., John Wiley & Sons, New York, NY, 1991, which is incorporated herein by reference. The proper selection of protecting groups for a particular synthesis is governed by the overall methods employed in the synthesis. For example, in "light-directed" synthesis, discussed herein, the protecting groups are photolabile protecting groups such as NVOC, MeNPoc, and those disclosed in co-pending Application PCT/US93/10162 (filed Oct. 22, 1993), incorporated herein by reference. In other methods, protecting groups are removed by chemical methods and include groups such as FMOC, DMT and others known to those of skill in the art.

A "purine" is a generic term based upon the specific compound "purine" having a skeletal structure derived from the fusion of a pyrimidine ring and an imidazole ring. It is generally, and herein, used to describe a generic class of compounds which have an atom or a group of atoms added to the parent purine compound, such as the bases found in the naturally occurring nucleic acids adenine (6-aminopurine) and guanine (2-amino-6-oxopurine), or less commonly occurring molecules such as 2-amino-adenine, N^6-methyladenine, or 2-methylguanine.

A "purine analogue" has a heterocyclic ring with structural similarities to a purine, in which an atom or group of atoms is substituted for an atom in the purine ring. For instance, in one embodiment, one or more N atoms of the purine heterocyclic ring are replaced by C atoms.

A "pyrimidine" is a compound with a specific heterocyclic diazine ring structure, but is used generically by persons of skill and herein to refer to any compound having a 1,3-diazine ring with minor additions, such as the common nucleic acid bases cytosine, thymine, uracil, 5-methylcytosine and 5-hydroxymethylcytosine, or the non-naturally occurring 5-bromo-uracil.

A "pyrimidine analogue" is a compound with structural similarity to a pyrimidine, in which one or more atom in the pyrimidine ring is substituted. For instance, in one embodiment, one or more of the N atoms of the ring are substituted with C atoms.

A "solid substrate" has fixed organizational support matrix, such as silica, polymeric materials, or glass. In some embodiments, at least one surface of the substrate is partially planar. In other embodiments it is desirable to physically separate regions of the substrate to delineate synthetic regions, for example with trenches, grooves, wells or the like. Example of solid substrates include slides, beads and arrays.

DESCRIPTION OF THE DRAWINGS

FIG. 1 shows four panels (FIG. 1A, FIG. 1B, FIG. 1C and FIG. 1D). FIGS. 1A and 1B graphically display the difference in fluorescence intensity between the matched and mismatched DNA probes. FIGS. 1C and 1D illustrate the difference in fluorescence intensity verses location on an example chip for DNA and RNA targets, respectively.

FIG. 2 is a graphic illustration of specific light-directed chemical coupling of oligonucleotide analogue monomers to an array.

FIG. 3 shows the relative efficiency and specificity of hybridization for immobilized probe arrays containing adenine versus probe arrays containing 2,6-diaminopurine nucleotides. (3'-CATCGTAGAA-5' (SEQ ID NO:1)).

FIG. 4 shows the effect of substituting adenine with 2,6-diaminopurine (D) in immobilized poly-dA probe arrays. (AAAAANAAAAA (SEQ ID NO:2)).

FIG. 5 shows the effects of substituting 5-propynyl-2'-deoxyuridine and 2-amino-2' deoxyadenosine in AT arrays on hybridization to a target nucleic acid. (ATATAATATA (SEQ ID NO:3) and CGCGCCGCGC (SEQ ID NO:4)).

FIG. 6 shows the effects of dI and 7-deaza-dG substitutions in oligonucleotide arrays. (3'-ATGTT(G1G2G3G4G5) CGGGT-5' (SEQ ID NO:5)).

DETAILED DESCRIPTION

Methods of synthesizing desired single stranded oligonucleotide and oligonucleotide analogue sequences are known to those of skill in the art. In particular, methods of synthesizing oligonucleotides and oligonucleotide analogues are found in, for example, *Oligonucleotide Synthesis: A Practical Approach*, Gait, ed., IRL Press, Oxford (1984); W. H. A. Kuijpers *Nucleic Acids Research* 18(17), 5197 (1994); K. L. Dueholm *J. Org. Chem.* 59, 5767–5773 (1994), and S. Agrawal (ed.) *Methods in Molecular Biology*, volume 20, each of which is incorporated herein by reference in its entirety for all purposes. Synthesizing unimolecular double-stranded DNA in solution has also been described. See, copending application Ser. No. 08/327,687, now U.S. Pat. No. 5,556,752 which is incorporated herein for all purposes.

Improved methods of forming large arrays of oligonucleotides, peptides and other polymer sequences with a minimal number of synthetic steps are known. See, Pirrung et al., U.S. Pat. No. 5,143,854 (see also, PCT Application No. WO 90/15070) and Fodor et al., PCT Publication No. WO 92/10092, which are incorporated herein by reference, which disclose methods of forming vast arrays of peptides, oligonucleotides and other molecules using, for example, light-directed synthesis techniques. See also, Fodor et al., (1991) *Science*, 251, 767–77 which is incorporated herein by reference for all purposes. These procedures for synthesis of polymer arrays are now referred to as VLSIPS™ procedures.

Using the VLSIP™ approach, one heterogenous array of polymers is converted, through simultaneous coupling at a number of reaction sites, into a different heterogenous array. See, U.S. application Ser. No. 07/796,243 now U.S. Pat. No. 5,384,261 and U.S. application Ser. No. 07/980,523 now U.S. Pat. No. 5,677,195, the disclosures of which are incorporated herein for all purposes.

The development of VLSIPS™ technology as described in the above-noted U.S. Pat. No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092 is considered pioneering technology in the fields of combinatorial synthesis and screening of combinatorial libraries. More recently, patent application Ser. No. 08/082,937, filed Jun. 25, 1993 (incorporated herein by reference), describes methods for making arrays of oligonucleotide probes that are used to check or determine a partial or complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific oligonucleotide sequence.

Combinatorial Synthesis of Oligonucleotide Arrays

VLSIPS™ technology provides for the combinatorial synthesis of oligonucleotide arrays. The combinatorial VLSIPS™ strategy allows for the synthesis of arrays containing a large number of related probes using a minimal number of synthetic steps. For instance, it is possible to synthesize and attach all possible DNA 8mer oligonucleotides (4^8, or 65,536 possible combinations) using only 32 chemical synthetic steps. In general, VLSIPS™ procedures provide a method of producing 4^n different oligonucleotide probes on an array using only 4n synthetic steps.

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, e.g., a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithogaphic mask is used selectively to expose functional groups which are then ready to react with incoming 5'-photoprotected nucleoside phosphoramidites. See, FIG. 2. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid surface. Combinatorial synthesis of different oligonucleotide analogues at different locations on the array is determined by the pattern of illumination during synthesis and the order of addition of coupling reagents.

In the event that an oligonucleotide analogue with a polyamide backbone is used in the VLSIPS™ procedure, it is generally inappropriate to use phosphoramidite chemistry to perform the synthetic steps, since the monomers do not attach to one another via a phosphate linkage. Instead, peptide synthetic method are substituted. See, e.g., Pirrung et al. U.S. Pat. No. 5,143,854.

Peptide nucleic acids are commercially available from, e.g., Biosearch, Inc. (Bedford, Mass.) which comprise a polyamide backbone and the bases found in naturally occurring nucleosides. Peptide nucleic acids are capable of binding to nucleic acids with high specificity, and are considered "oligonucleotide analogues" for purposes of this disclosure. Note that peptide nucleic acids optionally comprise bases other than those which are naturally occurring.

Hybridization of Nucleotide Analogues

The stability of duplexes formed between RNAs or DNAs are generally in the order of RNA:RNA>RNA:DNA>DNA:DNA, in solution. Long probes have better duplex stability with a target, but poorer mismatch discrimination than shorter probes (mismatch discrimination refers to the measured hybridization signal ratio between a perfect match probe and a single base mismatch probe. Shorter probes (e.g., 8-mers) discriminate mismatches very well, but the overall duplex stability is low. In order to optimize mismatch discrimination and duplex stability, the present invention provides a variety of nucleotide analogues incorporated into polymers and attached in an array to a solid substrate.

Altering the thermal stability (T_m) of the duplex formed between the target and the probe using, e.g., known oligonucleotide analogues allows for optimization of duplex stability and mismatch discrimination. One useful aspect of altering the T_m arises from the fact that Adenine-Thymine (A-T) duplexes have a lower T_m than Guanine-Cytosine (G-C) duplexes, due in part to the fact that the A-T duplexes have 2 hydrogen bonds per base-pair, while the G-C duplexes have 3 hydrogen bonds per base pair. In heterogeneous oligonucleotide arrays in which there is a non-uniform distribution of bases, it can be difficult to optimize hybridization conditions for all probes simultaneously. Thus, in some embodiments, it is desirable to destabilize G-C-rich duplexes and/or to increase the stability of A-T-rich duplexes while maintaining the sequence specificity of hybridization. This is accomplished, e.g., by replacing one or more of the native nucleotides in the probe (or the target) with certain modified, non-standard nucleotides. Substitution of guanine residues with 7-deazaguanine, for example, will generally destabilize duplexes, whereas substituting adenine residues with 2,6-diaminopurine will enhance duplex stability. A variety of other modified bases are also incorporated into nucleic acids to enhance or decrease overall duplex stability while maintaining specificity of hybridization. The incorporation of 6-aza-pyrimidine analogs into oligonucleotide probes generally decreases their binding affinity for complementary nucleic acids. Many 5-substituted pyrimidines substantially increase the stability of hybrids in which they have been substituted in place of the native pyrimidines in the sequence. Examples include 5-bromo-, 5-methyl-, 5-propynyl-, 5-(imidazol-2-yl)-and 5-(thiazol-2-yl)-derivatives of cytosine and uracil.

Many modified nucleosides, nucleotides and various bases suitable for incorporation into nucleosides are commercially available from a variety of manufacturers, including the SIGMA chemical company (Saint Louis, Mo.), R&D systems (Minneapolis, Minn.), Pharmacia LKB Biotechnology (Piscataway, N.J.), CLONTECH Laboratories, Inc. (Palo Alto, Calif.), Chem Genes Corp., Aldrich Chemical Company (Milwaukee, Wis.), Glen Research, Inc., GIBCO BRL Life Technologies, Inc. (Gaithersberg, Md.), Fluka Chemica-Biochemika Analytika (Fluka Chemie AG, Buchs, Switzerland), Invitrogen, San Diego, Calif., and Applied Biosystems (Foster City, Calif.), as well as many other commercial sources known to one of skill. Methods of attaching bases to sugar moieties to form nucleosides are known. See, e.g., Lukevics and Zablocka (1991), *Nucleoside Synthesis: Organosilicon Methods* Ellis Horwood Limited Chichester, West Sussex, England and the references therein. Methods of phosphorylating nucleosides to form nucleotides, and of incorporating nucleotides into oligonucleotides are also known. See, e.g., Agrawal (ed) (1993) *Protocols for Oligonucleotides and Analogues, Synthesis and Properties*, Methods in Molecular Biology volume 20, Humana Press, Towota, N.J., and the references therein. See also, Crooke and Lebleu, and Sanghvi and Cook, and the references cited therein, both supra.

Groups are also linked to various positions on the nucleoside sugar ring or on the purine or pyrimidine rings which may stabilize the duplex by electrostatic interactions with the negatively charged phosphate backbone, or through hydrogen bonding interactions in the major and minor groves. For example, adenosine and guanosine nucleotides are optionally substituted at the N^2 position with an imidazolyl propyl group, increasing duplex stability. Universal base analogues such as 3-nitropyrrole and 5-nitroindole are optionally included in oligonucleotide probes to improve duplex stability through base stacking interactions.

Selecting the length of oligonucleotide probes is also an important consideration when optimizing hybridization specificity. In general, shorter probe sequences are more specific than longer ones, in that the occurrence of a single-base mismatch has a greater destabilizing effect on the hybrid duplex. However, as the overall thermodynamic stability of hybrids decreases with length, in some embodiments it is desirable to enhance duplex stability for short probes globally. Certain modifications of the sugar moiety in oligonucleotides provide useful stabilization, and these can be used to increase the affinity of probes for complementary nucleic acid sequences. For example, 2'-O-methyl-, 2'-O-propyl-, and 2'-O-allyl-oligoribonucleotides have higher binding affinities for complementary RNA sequences than their unmodified counterparts. Probes comprised of 2'-fluoro-2'-deoxyollgoribonucleotides also form more stable hybrids with RNA than do their unmodified counterparts.

Replacement or substitution of the internucleotide phosphodiester linkage in oligo- or poly-nucleotides is also used to either increase or decrease the affinity of probe-target interactions. For example, substituting phosphodiester linkages with phosphorothioate or phosphorodithioate linkages generally lowers duplex stability, without affecting sequence specificity. Substitutions with a non-ionic methylphosphonate linkage (racemic, or preferably, Rp stereochemistry) have a stabilizing influence on hybrid formation. Neutral or cationic phosphoramidate linkages also result in enhanced duplex stabilization. The phosphate diester backbone has been replaced with a variety of other stabilizing, non-natural linkages which have been studied as potential antisense therapeutic agents. See, e.g., Crooke and Lebleu (eds) (1993) *Antisense Research Applications* CRC Press; and, Sanghvi and Cook (eds) (1994) *Carbohydrate modifications in Antisense Research* ACS Symp. Ser. #580 ACS, Washington DC. Very stable hybrids are formed between nucleic acids and probes comprised of peptide nucleic acids, in which the entire sugar-phosphate backbone has been replaced with a polyamide structure.

Another important factor which sometimes affects the use of oligonucleotide probe arrays is the nature of the target nucleic acid. Oligodeoxynucleotide probes can hybridize to DNA and RNA targets with different affinity and specificity. For example, probe sequences containing long "runs" of consecutive deoxyadenosine residues form less stable hybrids with complementary RNA sequences than with the complementary DNA sequences. Substitution of dA in the probe with either 2,6-diaminopurine deoxyriboside, or 2'-alkoxy- or 2'-fluoro-dA enhances hybridization with RNA targets.

Internal structure within nucleic acid probes or the targets also influences hybridization efficiency. For example, GC-rich sequences, and sequences containing "runs" of consecutive G residues frequently self-associate to form higher-order structures, and this can inhibit their binding to complementary sequences. See, Zimmermann et al. (1975) *J. Mol Biol* 92: 181; Kim (1991) *Nature* 351: 331; Sen and Gilbert (1988) *Nature* 335: 364; and Sunquist and Klug (1989) *Nature* 342: 825. These structures are selectively destabilized by the substitution of one or more guanine residues with one or more of the following purines or purine analogs: 7-deazaguanine, 8-aza-7-deazaguanine, 2-aminopurine, 1H-purine, and hypoxanthine, in order to enhance hybridization.

Modified nucleic acids and nucleic acid analogs can also be used to improve the chemical stability of probe arrays. For example, certain processes and conditions that are useful for either the fabrication or subsequent use of the arrays, may not be compatible with standard oligonucleotide chemistry, and alternate chemistry can be employed to overcome these problems. For example, exposure to acidic conditions will cause depurination of purine nucleotides, ultimately resulting in chain cleavage and overall degradation of the probe array. In this case, adenine and guanine are replaced with 7-deazaadenine and 7-deazaguanine, respectively, in order to stabilize the oligonucleotide probes towards acidic conditions which are used during the manufacture or use of the arrays.

Base, phosphate and sugar modifications are used in combination to make highly modified oligonucleotide analogues which take advantage of the properties of each of the various modifications. For example, oligonucleotides which have higher binding affinities for complementary sequences than their unmodified counterparts (e.g., 2'-O-methyl-, 2'-O-propyl-, and 2'-O-allyl oligonucleotides) can be incorporated into oligonucleotides with modified bases (deazaguanine, 8-aza-7-deazaguanine, 2-aminopurine, 1H-purine, hypoxanthine and the like) with non-ionic methylphosphonate linkages or neutral or cationic phosphoramidate linkages, resulting in additive stabilization of duplex formation between the oligonucleotide and a target nucleic acid. For instance, one preferred oligonucleotide comprises a 2'-O-methyl-2,6-diaminopurineriboside phosphorothioate. Similarly, any of the modified bases described herein can be incorporated into peptide nucleic acids, in which the entire sugar-phosphate backbone has been replaced with a polyamide structure.

Thermal equilibrium studies, kinetic "on-rate" studies, and sequence specificity analysis is optionally performed for any target oligonucleotide and probe or probe analogue. The data obtained shows the behavior of the analogues upon duplex formation with target oligonucleotides. Altered duplex stability conferred by using oligonucleotide analogue probes are ascertained by following, e.g., fluorescence signal intensity of oligonucleotide analogue arrays hybridized with a target oligonucleotide over time. The data allow optimization of specific hybridization conditions at, e.g., room temperature (for simplified diagnostic applications).

Another way of verifying altered duplex stability is by following the signal intensity generated upon hybridization with time. Previous experiments using DNA targets and DNA chips have shown that signal intensity increases with time, and that the more stable duplexes generate higher signal intensities faster than less stable duplexes. The signals reach a plateau or "saturate" after a certain amount of time due to all of the binding sites becoming occupied. These data allow for optimization of hybridization, and determination of equilibration conditions at a specified temperature.

Graphs of signal intensity and base mismatch positions are plotted and the ratios of perfect match versus mismatches calculated. This calculation shows the sequence specific properties of nucleotide analogues as probes. Perfect match/mismatch ratios greater than 4 are often desirable in an oligonucleotide diagnostic assay because, for a diploid genome, ratios of 2 have to be distinguished (e.g., in the case of a heterozygous trait or sequence).

Target Nucleic Acids Which Comprise Nucleotide Analogues

Modified nucleotides and nucleotide analogues are incorporated synthetically or enzymatically into DNA or RNA target nucleic acids for hybridization analysis to oligonucleotide arrays. The incorporation of nucleotide analogues in the target optimizes the hybridization of the target in terms of sequence specificity and/or the overall affinity of binding to oligonucleotide and oligonucleotide analogue probe arrays. The use of nucleotide analogues in either the oligonucleotide array or the target nucleic acid, or both, improves optimizability of hybridization interactions. Examples of useful nucleotide analogues which are substituted for naturally occurring nucleotides include 7-deazaguanosine, 2,6-diaminopurine nucleotides, 5-propynyl and other 5-substituted pyrimidine nucleotides, 2'-fluro and 2'-methoxy -2'-deoxynucleotides and the like.

These nucleotide analogues are incorporated into nucleic acids using the synthetic methods described supra, or using DNA or RNA polymerases. The nucleotide analogues are preferably incorporated into target nucleic acids using in vitro amplification methods such as PCR, LCR, Qβ-replicase expansion, in vitro transcription (e.g., nick translation or random-primer transcription) and the like. Alternatively, the nucleotide analogues are optionally incorporated into cloned nucleic acids by culturing a cell which comprises the cloned nucleic acid in media which includes a nucleotide analogue.

Similar to the use of nucleotide analogues in probe arrays, 7-deazaguanosine is used in target nucleic acids to substitute for G/dG to enhance target hybridization by reducing secondary structure in sequences containing runs of poly-G/dG. 6diaminopurine nucleotides substitute for A/dA to enhance target hybridization through enhanced H-bonding to T or U rich probes. 5-propynyl and other 5-substituted pyrimidine nucleotides substitute for natural pyrimidines to enhance target hybridization to certain purine rich probes. 2'-fluro and 2'-methoxy-2'-deoxynucleotides substitute for natural nucleotides to enhance target hybridization to similarly substituted probe sequences.

Synthesis of 5'-photoprotected 2'-O alkyl ribonucleotide analogues

The light-directed synthesis of complex arrays of nucleotide analogues on a glass surface is achieved by derivatizing cyanoethyl phosphoramidite nucleotides and nucleotide analogues (e.g., nucleoside analogues of uridine, thymidine, cytidine, adenosine and guanosine, with phosphates) with, for example, the photolabile MeNPoc group in the 5'-hydroxyl position instead of the usual dimethoxytrityl group. See, application SN PCT/US94/12305.

Specific base-protected 2'-O alkyl nucleosides are commercially available, from, e.g., Chem Genes Corp. (MA). The photolabile MeNPoc group is added to the 5'-hydroxyl position followed by phosphitylation to yield cyanoethyl phosphoramidite monomers. Commercially available nucleosides are optionally modified (e.g., by 2-O-alkylation) to create nucleoside analogues which are used to generate oligonucleotide analogues.

Modifications to the above procedures are used in some embodiments to avoid significant addition of MenPoc to the 3'-hydroxyl position. For instance, in one embodiment, a 2'-O-methyl ribonucleotide analogue is reacted with DMT-Cl {di(p-methoxyphenyl)phenylchloride} in the presence of pyridine to generate a 2'-O-methyl-5'-O-DMT ribonucleotide analogue. This allows for the addition of TBDMS to the 3'-O of the ribonucleoside analogue by reaction with TBDMS-Triflate (t-butyldimethylsilyltrifluoromethanesulfonate) in the presence of triethylamine in THF (tetrahydrofuran) to yield a 2'-O-methyl-3'-O-TBDMS-5'-O-DMT ribonucleotide base analogue. This analogue is treated with TCAA (trichloroacetic acid) to cleave off the DMT group, leaving a reactive hydroxyl group at the 5' position. MeNPoc is then added to the oxygen of the 5' hydroxyl group using MenPoc-Cl in the presence of pyridine. The TBDMS group is then cleaved with F^- (e.g., NaF) to yield a ribonucleotide base analogue with a MeNPoc group attached to the 5' oxygen on the nucleotide analogue. If appropriate, this analogue is phosphitylated to yield a phosphoramidite for oligonucleotide analogue synthesis. Other nucleosides or nucleoside analogues are protected by similar procedures.

Synthesis of Oligonucleotide Analogue Arrays on Chips

Other than the use of photoremovable protecting groups, the nucleoside coupling chemistry used in VLSIPS™ technology for synthesizing oligonucleotides and oligonucleotide analogues on chips is similar to that used for oligonucleotide synthesis. The oligonucleotide is typically linked to the substrate via the 3'-hydroxyl group of the oligonucleotide and a functional group on the substrate which results in the formation of an ether, ester, carbamate or phosphate ester linkage. Nucleotide or oligonucleotide analogues are attached to the solid support via carbon-carbon bonds using, for example, supports having (poly)trifluorochloroethylene surfaces, or preferably, by siloxane bonds (using, for example, glass or silicon oxide as the solid support). Siloxane bonds with the surface of the support are formed in one embodiment via reactions of surface attaching portions bearing trichlorosilyl or trialkoxysilyl groups. The surface attaching groups have a site for attachment of the oligonucleotide analogue portion. For example, groups which are suitable for attachment include amines, hydroxyl, thiol, and carboxyl. Preferred surface attaching or derivitizing portions include aminoalkylsilanes and hydroxyalkylsilanes. In particularly preferred embodiments, the surface attaching portion of the oligonucleotide analogue is either bis(2-hydroxyethyl)-aminopropyltriethoxysilane, n-(3-triethoxysilylpropyl)-4-hydroxybutylamide, aminopropyltriethoxysilane or hydroxypropyltriethoxysilane.

The oligoribonucleotides generated by synthesis using ordinary ribonucleotides are usually base labile due to the presence of the 2'-hydroxyl group. 2'-O-methyloligoribonucleotides (2'-OMeORNs), analogues of RNA where the 2'-hydroxyl group is methylated, are DNAse and RNAse resistant, making them less base labile. Sproat, B. S., and Lamond, A. I. in *Oligonucleotides and Analogues: A Practical Approach*, edited by F. Eckstein, New York: IRL Press at Oxford University Press, 1991, pp. 49–86, incorporated herein by reference for all purposes, have reported the synthesis of mixed sequences of 2'-O-Methoxy-oligoribonucleotides (2'-O-MeORNs) using dimethoxytrityl phosphoramidite chemistry. These 2'-O-MeORNs display greater binding affinity for complementary nucleic acids than their unmodified counterparts.

Other embodiments of the invention provide mechanical means to generate oligonucleotide analogues. These techniques are discussed in co-pending application Ser. No. 07/796,243, filed Nov. 22, 1991, which is incorporated herein by reference in its entirety for all purposes. Essentially, oligonucleotide analogue reagents are directed over the surface of a substrate such that a predefined array of oligonucleotide analogues is created. For instance, a series of channels, grooves, or spots are formed on or adjacent to a substrate. Reagents are selectively flowed through or deposited in the channels, grooves, or spots, forming an array having different oligonucleotides and/or oligonucleotide analogues at selected locations on the substrate.

Detection of Hybridization

In one embodiment, hybridization is detected by labeling a target with, e.g., fluorescein or other known visualization agents and incubating the target with an array of oligonucleotide analogue probes. Upon duplex formation by the target with a probe in the array (or triplex formation in embodiments where the array comprises unimolecular double-stranded probes), the fluorescein label is excited by, e.g., an argon laser and detected by viewing the array, e.g., through a scanning confocal microscope.

Sequencing by hybridization

Current sequencing methodologies are highly reliant on complex procedures and require substantial manual effort. Conventional DNA sequencing technology is a laborious procedure requiring electrophoretic size separation of labeled DNA fragments. An alternative approach involves a hybridization strategy carried out by attaching target DNA to a surface. The target is interrogated with a set of oligonucleotide probes, one at a time (see, application SN PCT/US94/12305).

A preferred method of oligonucleotide probe array synthesis involves the use of light to direct the synthesis of oligonucleotide analogue probes in high-density, miniaturized arrays. Matrices of spatially-defined oligonucleotide analogue probe arrays were generated. The ability to use these arrays to identify complementary sequences was demonstrated by hybridizing fluorescent labeled oligonucleotides to the matrices produced.

Oligonucleotide analogue arrays are used, e.g., to study sequence specific hybridization of nucleic acids, or protein-nucleic acid interactions. Oligonucleotide analogue arrays are used to define the thermodynamic and kinetic rules governing the formation and stability of oligonucleotide and oligonucleotide analogue complexes.

Oligonucleotide analogue Probe Arrays and Libraries

The use of oligonucleotide analogues in probe arrays provides several benefits as compared to standard oligonucleotide arrays. For instance, as discussed supra, certain oligonucleotide analogues have enhanced hybridization characteristics to complementary nucleic acids as compared with oligonucleotides made of naturally occurring nucleotides. One primary benefit of enhanced hybridization characteristics is that oligonucleotide analogue probes are optionally shorter than corresponding probes which do not include nucleotide analogues.

Standard oligonucleotide probe arrays typically require fairly long probes (about 15–25 nucleotides) to achieve strong binding to target nucleic acids. The use of such long probes is disadvantageous for two reasons. First, the longer the probe, the more synthetic steps must be performed to make the probe and any probe array comprising the probe. This increases the cost of making the probes and arrays. Furthermore, as each synthetic step results in less than 100% coupling for every nucleotide, the quality of the probes degrades as they become longer. Secondly, short probes provide better mis-match discrimination for hybridization to a target nucleic acid. This is because a single base mismatch for a short probe-target hybridization is less destabilizing than a single mismatch for a long probe-target hybridization. Thus, it is harder to distinguish a single probe-target mismatch when the probe is a 20-mer than when the probe is an 8-mer. Accordingly, the use of short oligonucleotide analogue probes reduces costs and increases mismatch discrimination in probe arrays.

The enhanced hybridization characteristics of oligonucleotide analogues also allows for the creation of oligonucleotide analogue probe arrays where the probes in the arrays have substantial secondary structure. For instance, the oligonucleotide analogue probes are optionally configured to be fully or partially double stranded on the array. The probes are optionally complexed with complementary nucleic acids, or are optionally unimolecular oligonucleotides with self-complementary regions. Libraries of diverse double-stranded oligonucleotide analogue probes are used, for example, in screening studies to determine binding affinity of nucleic acid binding proteins, drugs, or oligonucleotides (e.g., to examine triple helix formation). Specific oligonucleotide analogues are known to be conducive to the formation of unusual secondary structure. See, Durland (1995) *Bioconjugate Chem.* 6: 278–282. General strategies for using unimolecular double-stranded oligonucleotides as probes and for library generation is described in application Ser. No 08/327,687, and similar strategies are applicable to oligonucleotide analogue probes.

In general, a solid support, which optionally has an attached spacer molecule is attached to the distal end of the oligonucleotide analogue probe. The probe is attached as a single unit, or synthesized on the support or spacer in a monomer by monomer approach using the VLSIPS™ or mechanical partitioning methods described supra. Where the oligonucleotide analogue arrays are fully double-stranded, oligonucleotides (or oligonucleotide analogues) complementary to the probes on the array are hybridized to the array.

In some embodiments, molecules other than oligonucleotides, such as proteins, dyes, co-factors, linkers and the like are incorporated into the oligonucleotide analogue probe, or attached to the distal end of the oligomer, e.g., as a spacing molecule, or as a probe or probe target. Flexible linkers are optionally used to separate complementary portions of the oligonucleotide analogue.

The present invention also contemplates the preparation of libraries of oligonucleotide analogues having bulges or loops in addition to complementary regions. Specific RNA bulges are often recognized by proteins (e.g., TAR RNA is recognized by the TAT protein of HIV). Accordingly, libraries of oligonucleotide analogue bulges or loops are useful in a number of diagnostic applications. The bulge or loop can be present in the oligonucleotide analogue or linker portions.

Unimolecular analogue probes can be configured in a variety of ways. In one embodiment, the unimolecular probes comprise linkers, for example, where the probe is arranged according to the formula Y—L^1—X^1—L^2—X^2, in which Y represents a solid support, X^1 and X^2 represent a pair of complementary oligonucleotides or oligonucleotide analogues, L^1 represents a bond or a spacer, and L^2 represents a linking group having sufficient length such that X^1 and X^2 form a double-stranded oligonucleotide. The general synthetic and conformational strategy used in generating the double-stranded unimolecular probes is similar to that described in co-pending application Ser. No. 08/327,687, except that any of the elements of the probe (L^1, X^1, L^2 and X^2) comprises a nucleotide or an oligonucleotide analogue. For instance, in one embodiment X^1 is an oligonucleotide analogue.

The oligonucleotide analogue probes are optionally arranged to present a variety of moieties. For example, structural components are optionally presented from the middle of a conformationally restricted oligonucleotide analogue probe. In these embodiments, the analogue probes generally have the structure—X^1—Z—X^2 wherein X^{11} and X^{12} are complementary oligonucleotide analogues and Z is a structural element presented away from the surface of the probe array. Z can include an agonist or antagonist for a cell membrane receptor, a toxin, venom, viral epitope, hormone, peptide, enzyme, cofactor, drug, protein, antibody or the like.

General tiling strategies for detection of a Polymorphism in a target oligonucleotide

In diagnostic applications, oligonucleotide analogue arrays (e.g., arrays on chips, slides or beads) are used to determine whether there are any differences between a reference sequence and a target oligonucleotide, e.g., whether an individual has a mutation or polymorphism in a known gene. As discussed supra, the oligonucleotide target is optionally a nucleic acid such as a PCR amplicon which comprises one or more nucleotide analogues. In one embodiment, arrays are designed to contain probes exhibiting complementarity to one or more selected reference sequence whose sequence is known. The arrays are used to read a target sequence comprising either the reference sequence itself or variants of that sequence. Any polynucleotide of known sequence is selected as a reference sequence. Reference sequences of interest include sequences known to include mutations or polymorphisms associated with phenotypic changes having clinical significance in human patients. For example, the CFTR gene and P53 gene in humans have been identified as the location of several mutations resulting in cystic fibrosis or cancer respectively. Other reference sequences of interest include those that serve to identify pathogenic microorganisms and/or are the site of mutations by which such microorganisms acquire drug resistance (e.g., the HIV reverse transcriptase gene for HIV resistance). Other reference sequences of interest include regions where polymorphic variations are known to occur (e.g., the D-loop region of mitochondrial DNA). These reference sequences also have utility for, e.g., forensic, cladistic, or epidemiological studies.

Other reference sequences of interest include those from the genome of pathogenic viruses (e.g., hepatitis (A, B, or C), herpes virus (e.g., VZV, HSV-1, HAV-6, HSV-II, CMV, and Epstein Barr virus), adenovirus, influenza virus, flaviviruses, echovirus, rhinovirus, coxsackie virus, cornovirus, respiratory syncytial virus, mumps virus, rotavirus, measles virus, rubella virus, parvovirus, vaccinia virus, HTLV virus, dengue virus, papillomavirus, molluscum virus, poliovirus, rabies virus, JC virus and arboviral encephalitis virus. Other reference sequences of interest are from genomes or episomes of pathogenic bacteria, particularly regions that confer drug resistance or allow phylogenic characterization of the host (e.g., 16S rRNA or corresponding DNA). For example, such bacteria include chlamydia, rickettsial bacteria, mycobacteria, staphylococci, treptocci, pneumonococci, meningococci and conococci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lymes disease bacteria. Other reference sequences of interest include those in which mutations result in the following autosomal recessive disorders: sickle cell anemia, β-thalassemia, phenylketonuria, galactosemia, Wilson's disease, hemochromatosis, severe combined immunodeficiency, alpha-1-antitrypsin deficiency, albinism, alkaptonuria, lysosomal storage diseases and Ehlers-Danlos syndrome. Other reference sequences of interest include those in which mutations result in X-linked recessive disorders: hemophilia, glucose-6-phosphate dehydrogenase, agammaglobulimenia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease and fragile X-syndrome. Other reference sequences of interest includes those in which mutations result in the following autosomal dominant disorders: familial hypercholesterolemia, polycystic kidney disease, Huntington's disease, hereditary spherocytosis, Marfan's syndrome, von Willebrand's disease, neurofibromatosis, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, myotonic dystrophy, muscular dystrophy, osteogenesis imperfecta, acute intermittent porphyria, and von Hippel-Lindau disease.

Although an array of oligonucleotide analogue probes is usually laid down in rows and columns for simplified data processing, such a physical arrangement of probes on the solid substrate is not essential. Provided that the spatial location of each probe in an array is known, the data from the probes is collected and processed to yield the sequence of a target irrespective of the physical arrangement of the probes on, e.g., a chip. In processing the data, the hybridization signals from the respective probes is assembled into any conceptual array desired for subsequent data reduction, whatever the physical arrangement of probes on the substrate.

In one embodiment, a basic tiling strategy provides an array of immobilized probes for analysis of a target oligonucleotide showing a high degree of sequence similarity to one or more selected reference oligonucleotide (e.g., detection of a point mutation in a target sequence). For instance, a first probe set comprises a plurality of probes exhibiting perfect complementarity with a selected reference oligonucleotide. The perfect complementarity usually exists throughout the length of the probe. However, probes having a segment or segments of perfect complementarity that is/are flanked by leading or trailing sequences lacking complementarity to the reference sequence can also be used. Within a segment of complementarity, each probe in the first probe set has at least one interrogation position that corresponds to a nucleotide in the reference sequence. The interrogation position is aligned with the corresponding nucleotide in the reference sequence when the probe and reference sequence are aligned to maximize complementarity between the two. If a probe has more than one interrogation position, each corresponds with a respective nucleotide in the reference sequence. The identity of an interrogation position and corresponding nucleotide in a particular probe in the first probe set cannot be determined simply by inspection of the probe in the first set. An interrogation position and corresponding nucleotide is defined by the comparative structures of probes in the first probe set and corresponding probes from additional probe sets.

For each probe in the first set, there are, for purposes of the present illustration, multiple corresponding probes from additional probe sets. For instance, there are optionally probes corresponding to each nucleotide of interest in the reference sequence. Each of the corresponding probes has an interrogation position aligned with that nucleotide of interest. Usually, the probes from the additional probe sets are identical to the corresponding probe from the first probe set with one exception. The exception is that at the interrogation position, which occurs in the same position in each of the corresponding probes from the additional probe sets. This position is occupied by a different nucleotide in the corresponding probe sets. Other tiling strategies are also employed, depending on the information to be obtained.

The probes are oligonucleotide analogues which are capable of hybridizing with a target nucleic sequence by complementary base-pairing. Complementary base pairing includes sequence-specific base pairing, which comprises, e.g., Watson-Crick base pairing or other forms of base pairing such as Hoogsteen base pairing. The probes are attached by any appropriate linkage to a support. 3' attachment is more usual as this orientation is compatible with the preferred chemistry used in solid phase synthesis of oligonucleotides and oligonucleotide analogues (with the exception of, e.g., analogues which do not have a phosphate backbone, such as peptide nucleic acids).

EXAMPLES

The following examples are provided by way of illustration only and not by way of limitation. A variety of parameters can be changed or modified to yield essentially similar results.

One approach to enhancing oligonucleotide hybridization is to increase the thermal stability (T_m) of the duplex formed between the target and the probe using oligonucleotide analogues that are known to increase T_m's upon hybridization to DNA. Enhanced hybridization using oligonucleotide analogues is described in the examples below, including enhanced hybridization in oligonucleotide arrays.

Example 1

Solution oligonucleotide melting T_m

The T_m of 2'-O-methyl oligonucleotide analogues was compared to the T_m for the corresponding DNA and RNA sequences in solution. In addition, the T_m of 2'-O-methyl oligonucleotide:DNA, 2'-O-methyl oligonucleotide:RNA and RNA:DNA duplexes in solution was also determined.

The T_m was determined by varying the sample temperature and monitoring the absorbance of the sample solution at 260 nm. The oligonucleotide samples were dissolved in a 0.1M NaCl solution with an oligonucleotide concentration of 2 μM. Table 1 summarizes the results of the experiment. The results show that the hybridization of DNA in solution has approximately the same T_m as the hybridization of DNA with a 2'-O-methyl-substituted oligonucleotide analogue. The results also show that the T_m for the 2'-O-methyl-substituted oligonucleotide duplex is higher than that for the corresponding RNA:2'-O-methyl-substituted oligonucleotide duplex, which is higher than the T_m for the corresponding DNA:DNA or RNA:DNA duplex.

TABLE 1

Solution Oligonucleotide Melting Experiments
(+) = Target Sequence
(5'-CTGAACGGTAGCATCTTGAC-3')(SEQ ID NO: 6)*
(-) = Complementary Sequence
(5'GTCAAGATGCTACCGTTCAG-3')(SEQ ID NO: 7)*

Type of Oligonucleotide, Target Sequence (+)	Type of Oligonucleotide, Complementary Sequence (+)	T_m (° C.)
DNA(+)	DNA(-)	61.6
DNA(+)	2'OMe(-)	58.6
2'OMe(+)	DNA(-)	61.6
2'OMe(+)	2'OMe(-)	78.0
RNA(+)	DNA(-)	58.2
RNA(+)	2'OMe(-)	73.6

*T refers to thymine for the DNA oligonucleotides, or uracil for the RNA oligonucleotides.

Example 2

Array hybridization experiments with DNA chips and oligonucleotide analogue targets

A variable length DNA probe array on a chip was designed to discriminate single base mismatches in the 3 corresponding sequences 5'-CTGAACGGTAGCATCTTGAC-3' (SEQ ID NO:6) (DNA target), 5'-CUGAACGGUAGCAUCUUGAC-3' (SEQ ID NO:8) (RNA target) and 5'-CUGAACGGUAGCAUCUUGAC-3' (SEQ ID NO:9) (2'-O-methyl oligonucleotide target), and generated by the VLSIPS™ procedure. The Chip was designed with adjacent 12-mers and 8-mers which overlapped with the 3 target sequences as shown in Table 2.

TABLE 2

Array hybridization Experiments	
Target 1 (DNA)	5'-CTGAACGGTAGCATCTTGAC-3' (SEQ ID NO: 6)
8-mer probe (complement)	――――――
12-mer probe (complement)	―――――――――
Target 2 (RNA)	5'-CUGAACGGUAGCAUCUUGAC-3' (SEQ ID NO: 8)
8-mer probe (complement)	――――――
12-mer probe (complement)	―――――――――
Target 3 (2'-O-Me oligo)	5'-CUGAACGGUAGCAUCUUGAC-3' (SEQ ID NO: 9)
8-mer probe (complement)	――――――
12-mer probe (complement)	―――――――――

Target oligos were synthesized using standard techniques. The DNA and 2'-O-methyl oligonucleotide analogue target oligonucleotides were hybridized to the chip at a concentration of 10 nM in 5× SSPE at 20° C. in sequential experiments. Intensity measurements were taken at each probe position in the 8-mer and 12-mer arrays over time. The rate of increase in intensity was then plotted for each probe position. The rate of increase in intensity was similar for both targets in the 8-mer probe arrays, but the 12-mer probes hybridized more rapidly to the DNA target oligonucleotide.

Plots of intensity versus probe position were generated for the RNA, DNA and 2-O-methyl oligonucleotides to ascertain mismatch discrimination. The 8-mer probes displayed similar mismatch discrimination against all targets. The 12-mer probes displayed the highest mismatch discrimination for the DNA targets, followed by the 2'-O-methyl target, with the RNA target showing the poorest mismatch discrimination.

Thermal equilibrium experiments were performed by hybridizing each of the targets to the chip for 90 minutes at 5° C. temperature intervals. The chip was hybridized with the target in 5× SSPE at a target concentration of 10 nM. Intensity measurements were taken at the end of the 90 minute hybridization at each temperature point as described above. All of the targets displayed similar stability, with minimal hybridization to the 8-mer probes at 30° C. In addition, all of the targets showed similar stability in hybridizing to the 12-mer probes. Thus, the 2'-O-methyl oligonucleotide target had similar hybridization characteristics to DNA and RNA targets when hybridized against DNA probes.

Example 3

2'-O-methyl-substituted oligonucleotide chips

DMT-protected DNA and 2'-O-methyl phosphoramidites were used to synthesize 8-mer probe arrays on a glass slide using the VLSIPS™ method. The resulting chip was hybridized to DNA and RNA targets in separate experiments. The target sequence, the sequences of the probes on the chip and the general physical layout of the chip is described in Table 3.

The chip was hybridized to the RNA and DNA targets in successive experiments. The hybridization conditions used were 10 nM target, in 5× SSPE. The chip and solution were heated from 20° C. to 50° C., with a fluorescence measurement taken at 5 degree intervals as described in SN PCT/US94/12305. The chip and solution were maintained at each temperature for 90 minutes prior to fluorescence measurements. The results of the experiment showed that DNA probes were equal or superior to 2'-O-methyl oligonucleotide analogue probes for hybridization to a DNA target, but that the 2'-O-methyl analogue oligonucleotide probes showed dramatically better hybridization to the RNA target than the DNA probes. In addition, the 2'-O-methyl analogue oligonucleotide probes showed superior mismatch discrimination of the RNA target compared to the DNA probes. The difference in fluorescence intensity between the matched and mismatched analogue probes was greater than the difference between the matched and mismatched DNA probes, dramatically increasing the signal-to-noise ratio. FIG. 1 displays the results graphically (FIGS. 1A and 1B). (M) and (P) indicate mismatched and perfectly matched probes, respectively. (FIGS. 1C and 1D) illustrates the fluorescence intensity versus location on an example chip for the various probes at 20° C. using DNA and RNA targets, respectively.

TABLE 3

2'-O-methyl Oligonucleotide Analogues on a Chip.	
Target Sequence (DNA):	5'-CTGAACGGTAGCATCTTGAC-3' (SEQ ID NO: 6)
Target Sequence (RNA):	5'-CUGAACGGUAGCAUCUUGAC-3' (SEQ ID NO: 8)
Matching DNA oligonucleotide probe {DNA (M)}	5'-CTTGCCAT (SEQ ID NO: 10)
Matching 2'-O-methyl oligonucleotide analogue probe {2'OMe (M)}	5'-CUUGCCAU (SEQ ID NO: 11)
DNA oligonucleotide probe with 1 base mismatch {DNA (P)}	5'-CTTGCTAT (SEQ ID NO: 12)
2'-O-methyl oligonucleotide analogue probe with 1 base mismatch {2'OMe (M)}	5'-CUUGCUAU (SEQ ID NO: 13)
SCHEMATIC REPRESENTATION OF 2'-O-METHYL/DNA CHIP	
Matching 2'-O-methyl oligonucleotide analogue probe	
2'-O-methyl oligonucleotide analogue probe with 1 base mismatch	
DNA oligonucleotide probe with 1 base mismatch	
Matching DNA oligonucleotide probe	

Example 4

Synthesis of oligonucleotide analogues

The reagent MeNPoc-Cl group reacts non-selectively with both the 5' and 3' hydroxyls on 2'-O-methyl nucleoside analogues. Thus, to generate high yields of 5'-O-MeNPoc-2'-O-methylribonucleoside analogues for use in oligonucleotide analogue synthesis, the following protection-deprotection scheme was utilized.

The protective group DMT was added to the 5'-O position of the 2'-O-methylribonucleoside analogue in the presence of pyridine. The resulting 5'-O-DMT protected analogue was reacted with TBDMS-Triflate in THF, resulting in the addition of the TBDMS group to the 3'-O of the analogue. The 5'-DMT group was then removed with TCAA to yield a free OH group at the 5' position of the 2'-O-methyl ribonucleoside analogue, followed by the addition of MeNPoc-Cl in the presence of pyridine, to yield 5'-O-MeNPoc-3'-O-TBDMS-2'-O-methyl ribonucleoside analogue. The TBDMS group was then removed by reaction with NaF, and the 3'-OH group was phosphitylated using standard techniques.

Two other potential strategies did not result in high specific yields of 5'-O-MeNPoc-2'-O-methylribonucleoside. In the first, a less reactive MeNPoc derivative was synthesized by reacting MeNPoc-Cl with N-hydroxy succimide to yield MeNPoc-NHS. This less reactive photocleavable group (MeNPoc-NHS) was found to react exclusively with the 3' hydroxyl on the 2'-O-methylribonucleoside analogue. In the second strategy, an organotin protection scheme was used. Dibutyltin oxide was reacted with the 2'-O-methylribonucleoside analogue followed by reaction with MeNPoc. Both 5'-O-MeNPoc and 3'-O-MeNPoc 2'-O-methylribonucleoside analogues were obtained.

Example 5

Hybridization to mixed-sequence oligodeoxynucleotide probes substituted with 2-amino-2'-deoxyadenosine (D)

To test the effect of a 2-amino-2'-deoxyadenosine (D) substitution in a heterogeneous probe sequence, two 4×4 oligodeoxynucleotide arrays were constructed using VLSIPS™ methodology and 5'-O-MeNPOC-protected deoxynucleoside phosphoramidites. Each array was comprised of the following set of probes based on the sequence (3')-CATCGTAGAA-(5') (SEQ ID NO:1):

1.-(HEG)-(3')-CATN$_1$GTAGAA-(5') (SEQ ID NO:14)

2.-(HEG)-(3')-CATCN$_2$TAGAA-(5') (SEQ ID NO:15)

3.-(HEG)-(3')-CATCGN$_3$AGAA-(5') (SEQ ID NO:16)

4.-(HEG)-(3')-CATCGTN$_4$GAA-(5') (SEQ ID NO:17)

where HEG=hexaethyleneglycol linker, and N is either A,G,C or T, so that probes are obtained which contain single mismatches introduced at each of four central locations in the sequence. The first probe array was constructed with all natural bases. In the second array, 2-amino-2'-deoxyadenosine (D) was used in place of adenosine (A). Both arrays were hybridized with a 5'-fluorescein-labeled oligodeoxynucleotide target, (5')-Fl-d(CTGAACGGTAGCATCTTGAC)-(3') (SEQ ID NO:18), which contained a sequence (in bold) complementary to the base probe sequence. The hybridization conditions were: 10 nM target in 5× SSPE buffer at 22° C. with agitation. After 30 minutes, the chip was mounted on the flowcell of a scanning laser confocal fluorescence microscope, rinsed briefly with 5× SSPE buffer at 22° C., and then a surface fluorescence image was obtained.

The relative efficiency of hybridization of the target to the complementary and single-base mismatched probes was determined by comparing the average bound surface fluorescence intensity in those regions of the of the array containing the individual probe sequences. The results (FIG. 3) show that a 2-amino-2'-deoxyadenosine (D) substitution in a heterogeneous probe sequence is a relatively neutral one, with little effect on either the signal intensity or the specificity of DNA-DNA hybridization, under conditions where the target is in excess and the probes are saturated.

Example 6

Hybridization to a dA-homopolymer oligodeoxynucleotide probe substituted with 2-amino-2'-deoxyadenosine (D)

The following experiment was performed to compare the hybridization of 2'-deoxyadenosine containing homopolymer arrays with 2-amino-2'-deoxyadenosine homopolymer arrays. The experiment was performed on two 11-mer oligodeoxynucleotide probe containing arrays. Two 11-mer oligodeoxynucleotide probe sequences were synthesized on a chip using 5'-O-MeNPOC-protected nucleoside phosphoramidites and standard VLSIPS™ methodology.

The sequence of the first probe was: (HEG)-(3')-d(AAAAANAAAAA)-(5') (SEQ ID NO:19); where HEG=hexaethyleneglycol linker, and N is either A,G,C or T. The second probe was the same, except that dA was replaced by 2-amino-2'-deoxyadenosine (D). The chip was hybridized with a 5'-fluorescein-labeled oligodeoxynucleotide target, (5')-Fl-d(TTTTTGTTTTT)-(3') (SEQ ID NO:20), which contained a sequence complementary to the probe sequences where N=C. Hybridization conditions were 10 nM target in 5× SSPE buffer at 22° C. with agitation. After 15 minutes, the chip was mounted on the flowcell of a scanning laser confocal fluorescence microscope, rinsed briefly with 5× SSPE buffer at 22° C. (low stringency), and a surface fluorescence image was obtained. Hybridization to the chip was continued for another 5 hours, and a surface fluorescence image was acquired again. Finally, the chip was washed briefly with 0.5× SSPE (high-stringency), then with 5× SSPE, and re-scanned.

The relative efficiency of hybridization of the target to the complementary and single-base mismatched probes was determined by comparing the average bound surface fluorescence intensity in those regions of the of the array containing the individual probe sequences. The results (FIG. 4) indicate that substituting 2'-deoxyadenosine with 2-amino-2'-deoxyadenosine in a d(A)$_n$ homopolymer probe sequence results in a significant enhancement in specific hybridization to a complementary oligodeoxynucleotide sequence.

Example 7

Hybridization to alternating A-T oligodeoxynucleotide probes substituted with 5-propynyl-2'-deoxyuridine (P) and 2-amino-2'-deoxyadenosine (D)

Commercially available 5'-DMT-protected 2'-deoxynucleoside/nucleoside-analog phosphoramidites (Glen Research) were used to synthesize two decanucleotide probe sequences on separate areas on a chip using a modified VLSIPS™ procedure. In this procedure, a glass substrate is initially modified with a terminal-MeNPOC-protected hexaethyleneglycol linker. The substrate was exposed to light through a mask to remove the protecting group from the linker in a checkerboard pattern. The first probe sequence was then synthesized in the exposed region using DMT-phosphoramidites with acid-deprotection cycles, and the sequence was finally capped with $(MeO)_2PNiPr_2$/tetrazole followed by oxidation. A second checkerboard exposure in a different (previously unexposed) region of the chip was then performed, and the second probe sequence was synthesized by the same procedure. The sequence of the first "control" probe was: -(HEG)-(3')-CGCGCCGCGC-(5') (SEQ ID NO:21); and the sequence of the second probe was one of the following:

1.-(HEG)-(3')-d(ATATAATATA)-(5') (SEQ ID NO:22)
2.-(HEG)-(3')-d(APAPAAPAPA)-(5') (SEQ ID NO:23)
3.-(HEG)-(3')-d(DTDTDDTDTD)-(5') (SEQ ID NO:24)
4.-(HEG)-(3')-d(DPDPDDPDPD)-(5') (SEQ ID NO:25)

where HEG=hexaethyleneglycol linker, A=2'-deoxyadenosine, T=thymidine, D=2-amino-2'-deoxyadenosine, and P=5-propynyl-2'-deoxyuridine. Each chip was then hybridized in a solution of a fluorescein-labeled oligodeoxynucleotide target, (5')-Fluorescein-d (TATATTATAT)-(HEG)-d(GCGCGGCGCG)-(3') (SEQ ID NO:26 and SEQ ID NO:27), which is complementary to both the A/T and G/C probes. The hybridization conditions were: 10 nM target in 5× SSPE buffer at 22° C. with gentle shaking. After 3 hours, the chip was mounted on the flowcell of a scanning laser confocal fluorescence microscope, rinsed briefly with 5× SSPE buffer at 22° C., and then a surface fluorescence image was obtained. Hybridization to the chip was continued overnight (total hybridization time=20hr), and a surface fluorescence image was acquired again.

The relative efficiency of hybridization of the target to the A/T and substituted A/T probes was determined by comparing the average surface fluorescence intensity bound to those parts of the chip containing the A/T or substituted probe to the fluorescence intensity bound to the G/C control probe sequence. The results (FIG. 5) show that 5-propynyl-dU and 2-amino-dA substitution in an A/T-rich probe significantly enhances the affinity of an oligonucleotide analogue for complementary target sequences. The unsubstituted A/T-probe bound only 20% as much target as the all-G/C-probe of the same length, while the D- & P-substituted A/T probe bound nearly as much (90%) as the G/C-probe. Moreover, the kinetics of hybridization are such that, at early times, the amount of target bound to the substituted A/T probes exceeds that which is bound to the all-G/C probe.

Example 8

Hybridization to oligodeoxynucleotide probes substituted with 7-deaza-2'-deoxyguanosine (ddG) and 2'-deoxyinosine (dI)

A 16×64 oligonucleotide array was constructed using VLSIPS™ methodology, with 5'-O-MeNPOC-protected nucleoside phosphoramidites, including the analogs ddG, and dI. The array was comprised of the set of probes represented by the following sequence: -(linker)-(3')-d(A T G T T G_1 G_2 G_3 G_4 G_5 C G G G T)-(5'); (SEQ ID NO:28) where underlined bases are fixed, and the five internal deoxyguanosines (G_{1-5}) are substituted with G, ddG, dI, and T in all possible (1024 total) combinations. A complementary oligonucleotide target, labeled with fluorescein at the 5'-end: (5')-Fl-d(C A A T A C A A C C C C C G C C C A T C C)-(3') (SEQ ID NO:29), was hybridized to the array. The hybridization conditions were: 5 nM target in 6× SSPE buffer at 22° C. with shaking. After 30 minutes, the chip was mounted on the flowcell of an Affymetrix scanning laser confocal fluorescence microscope, rinsed once with 0.25 × SSPE buffer at 22° C., and then a surface fluorescence image was acquired.

The "efficiency" of target hybridization to each probe in the array is proportional to the bound surface fluorescence intensity in the region of the chip where the probe was synthesized. The relative values for a subset of probes (those containing dG→ddG and dG→dI substitutions only) are shown in FIG. 6. Substitution of guanosine with 7-deazaguanosine within the internal run of five G's results in a significant enhancement in the fluorescence signal intensity which measures hybridization. Deoxyinosine substitutions also enhance hybridization to the probe, but to a lesser extent. In this example, the best overall enhancement is realized when the dG "run" is ~40–60% substituted with 7-deaza-dG, with the substitutions distributed evenly throughout the run (i.e., alternating dG/deaza-dG).

Example 9

Synthesis of 5'-MeNPOC-2'-deoxyinosine-3'-(N,N-diisopropyl-2-cyanoethyl)phosphoramidite

2'-deoxyinosine (5.0 g, 20 mmole) was dissolved in 50 ml of dry DMF, and 100 ml dry pyridine was added and evaporated three times to dry the solution. Another 50ml pyridine was added, the solution was cooled to −20° C. under argon, and 13.8 g (50 mmole) of MeNPOC-chloride in 20 ml dry DCM was then added dropwise with stirring over 60 minutes. After 60 minutes, the cold bath was removed, and the solution was allowed to stir overnight at room temperature. Pyridine and DCM were removed by evaporation, 500 ml of ethyl acetate was added, and the solution was washed twice with water and then with brine (200 ml each). The aqueous washes were combined and back-extracted twice with ethyl acetate, and then all of the organic layers were combined, dried with Na_2SO_4, and evaporated under vacuum. The product was recrystallized from DCM to obtain 5.0 g (50% yield) of pure 5'-O-MeNPOC-2'-deoxyinosine as a yellow solid (99% purity, according to ^{1}H-NMR and HPLC analysis).

The MeNPOC-nucleoside (2.5 g, 5.1 mmole) was suspended in 60 ml of dry CH_3CN and phosphitylated with 2-cyanoethyl-N,N,N',N'-tetraisopropylphosphorodiamidite (1.65 g/1.66 ml; 5.5 mmole) and 0.47 g (2.7 mmole) of diisopropylammonium tetrazolide, according to the published procedure of Barone, et al. (*Nucleic Acids Res.* (1984) 12, 4051–61). The crude phosphoramidite was purified by flash chromatography on silica gel (90:8:2 DCM-MeOH-Et_3N), co-evaporated twice with anhydrous acetonitrile and dried under vacuum for ~24 hours to obtain 2.8 g (80%) of the pure product as a yellow solid (98% purity as determined by ^{1}H/^{31}P-NMR and HPLC).

Example 10

Synthesis of 5'-MeNPOC-7-deaza-2'-deoxy(N2-isobutyryl)-guanosine-3'-(N,N-diisopropyl-2-cyanoethyl) phosphoramidite.

The protected nucleoside 7-deaza-2'-deoxy(N2-isobutyryl)guanosine (1.0 g, 3 mmole; Chemgenes Corp., Waltham, Mass.) was dried by co-evaporating three times with 5 ml anhydrous pyridine and dissolved in 5 ml of dry pyridine-DCM (75:25 by vol.). The solution was cooled to −45° C. (dry ice/CH_3CN) under argon, and a solution of 0.9 g (3.3 mmole) MeNPOC-Cl in 2 ml dry DCM was then added dropwise with stirring. After 30 minutes, the cold bath was removed, and the solution allowed to stir overnight at room temperature. The solvents were evaporated, and the crude material was purified by flash chromatography on silica gel (2.5%-5% MeOH in DCM) to yield 1.5 g (88% yield) 5'-MeNPOC-7-deaza-2'-deoxy(N2-isobutyryl) guanosine as a yellow foam. The product was 98% pure according to ^{1}H-NMR and HPLC analysis.

The MeNPOC-nucleoside (1.25 g, 2.2 mmole) was phosphitylated according to the published procedure of Barone, et al. (*Nucleic Acids Res.* (1984) 12, 4051–61). The crude product was purified by flash chromatography on silica gel (60:35:5 hexane-ethyl acetate-Et_3N), co-evaporated twice with anhydrous acetonitrile and dried under vacuum for ~24 hours to obtain 1.3 g (75%) of the pure product as a yellow solid (98% purity as determined by $^1H/^{31}P$-NMR and HPLC).

Example 11

Synthesis of 5'-MeNPOC-2,6-bis(phenoxyacetyl) -2,6-diaminopurine-2'-deoxyriboside-3'-(N,N-diisopropyl-2-cyanoethyl)phosphoramidite.

The protected nucleoside 2,6-bis(phenoxyacetyl) -2,6-diaminopurine-2'-deoxyriboside (8 mmole, 4.2 g) was dried by coevaporating twice from anhydrous pyridine, dissolved in 2:1 pyridine/DCM (17.6 ml) and then cooled to −40° C. MeNPOC-chloride (8 mmole, 2.18 g) was dissolved in DCM (6.6 mls) and added to reaction mixture dropwise. The reaction was allowed to stir overnight with slow warming to room temperature. After the overnight stirring, another 2 mmole (0.6 g) in DCM (1.6 ml) was added to the reaction at −40° C. and stirred for an additional 6 hours or until no unreacted nucleoside was present. The reaction mixture was evaporated to dryness, and the residue was dissolved in ethyl acetate and washed with water twice, followed by a wash with saturated sodium chloride. The organic layer was dried with $MgSO_4$, and evaporated to a yellow solid which was purified by flash chromatography in DCM employing a methanol gradient to elute the desired product in 51% yield.

The 5'-MeNPOC-nucleoside (4.5 mmole, 3.5 g) was phosphitylated according to the published procedure of Barone, et al. (*Nucleic Acids Res.* (1984) 12, 4051–61). The crude product was purified by flash chromatography on silica gel (99:0.5:0.5 DCM-MeOH-Et_3N). The pooled fractions were evaporated to an oil, redissolved in a minimum amount of DCM, precipitated by the addition of 800 ml ice cold hexane, filtered, and then dried under vacuum for ~24 hours.

Overall yield was 56%, at greater than 96% purity by HPLC and $^1H/^{31}P$-NMR.

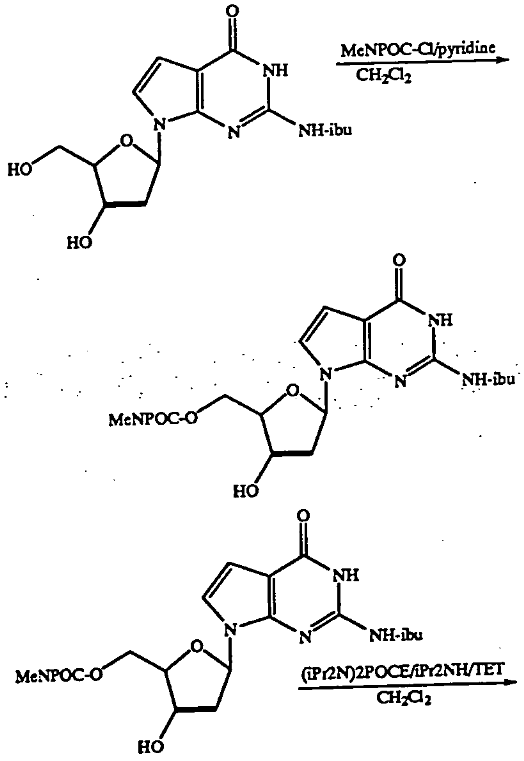
Example 12

5'-O-MeNPOC-protected phosphoramidites for incorporating 7-deaza-2'deoxyguanosine and 2'-deoxyinosine into VLSSIPS™ Oligonucleotide Arrays

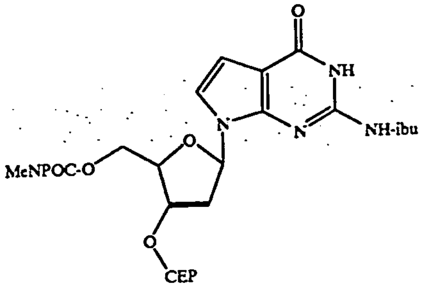
VLSIPS oligonucleotide probe arrays in which all or a subset of all guanosine residues are substitutes with 7-deaza-2'-deoxyguanosine and/or 2'-deoxyinosine are highly desirable. This is because guanine-rich regions of nucleic acids associate to form multi-stranded structures. For example, short tracts of G residues in RNA and DNA commonly associate to form tetrameric structures (Zimmermann et al. (1975) *J. Mol. Biol.* 92: 181; Kim, J. (1991) *Nature* 351: 331; Sen et al. (1988) *Nature* 335: 364; and Sunquist et al. (1989) *Nature* 342: 825). The problem this poses to chip hybridization-based assays is that such structures may compete or interfere with normal hybridization between complementary nucleic acid sequences. However, by substituting the 7-deaza-G analog into G-rich nucleic acid sequences, particularly at one or more positions within a run of G residues, the tendency for such probes to form higher-order structures is suppressed, while maintaining essentially the same affinity and sequence specificity in double-stranded structures. This has been exploited in order to reduce band compression in sequencing gels (Mizusawa, et al. (1986) N.A.R. 14: 1319) to improve target hybridization to G-rich probe sequences in VLSIPS arrays. Similar results are achieved using inosine (see also, Sanger et al. (1977) P.N.A.S. 74: 5463).

For facile incorporation of 7-deaza-2'-deoxyguanosine and 2'-deoxyinosine into oligonucleotide arrays using VLSIPS™ methods, a nucleoside phosphoramidite comprising the analogue base which has a 5'-O'-MeNPOC-protecting group is constructed. This building block was prepared from commercially available nucleosides according to Scheme I. These amidites pass the usual tests for coupling efficiency and photolysis rate.

SCHEME I



-continued



Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, modifications can be made thereto without departing from the spirit or scope of the appended claims.

All publications and patent applications cited in this application are herein incorporated by reference for all purposes as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.

SEQUENCE LISTING

```
(1) GENERAL INFORMATION:

   (iii) NUMBER OF SEQUENCES: 29


(2) INFORMATION FOR SEQ ID NO:1:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

AAGATGCTAC                                                          10


(2) INFORMATION FOR SEQ ID NO:2:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 11 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

AAAAANAAAA A                                                        11


(2) INFORMATION FOR SEQ ID NO:3:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

ATATAATATA                                                          10


(2) INFORMATION FOR SEQ ID NO:4:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear
```

-continued

```
    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

CGCGCCGCGC                                                         10


(2) INFORMATION FOR SEQ ID NO:5:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 15 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 6..10
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = guanosine (G),
               2',3'-dideoxyguanine (ddG),
               2'-deoxyinosine (dI) or thymine (T)"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

TGGGCNNNNN TTGTA                                                   15


(2) INFORMATION FOR SEQ ID NO:6:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 20 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..20
          (D) OTHER INFORMATION: /note= "Target DNA sequence"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CTGAACGGTA GCATCTTGAC                                              20


(2) INFORMATION FOR SEQ ID NO:7:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 20 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..20
          (D) OTHER INFORMATION: /note= "Complementary DNA sequence"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GTCAAGATGC TACCGTTCAG                                              20


(2) INFORMATION FOR SEQ ID NO:8:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 20 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear
```

-continued

(ii) MOLECULE TYPE: RNA

(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..20
(D) OTHER INFORMATION: /note= "Target RNA sequence"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

CUGAACGGUA GCAUCUUGAC 20

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "2'-O-methyl oligonucleotide"

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 1
(D) OTHER INFORMATION: /mod_base= cm

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 2
(D) OTHER INFORMATION: /mod_base= um

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 3
(D) OTHER INFORMATION: /mod_base= gm

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 4
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "2'-O-methyladenosine"

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 5
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "2'-O-methyladenosine"

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 6
(D) OTHER INFORMATION: /mod_base= cm

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 7
(D) OTHER INFORMATION: /mod_base= gm

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 8
(D) OTHER INFORMATION: /mod_base= gm

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 9
(D) OTHER INFORMATION: /mod_base= um

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 10
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "2'-O-methyladenosine"

-continued

```
     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 11
          (D) OTHER INFORMATION: /mod_base= gm

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 12
          (D) OTHER INFORMATION: /mod_base= cm

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 13
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "2'-O-methyladenosine"

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 14
          (D) OTHER INFORMATION: /mod_base= um

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 15
          (D) OTHER INFORMATION: /mod_base= cm

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 16
          (D) OTHER INFORMATION: /mod_base= um

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 17
          (D) OTHER INFORMATION: /mod_base= um

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 18
          (D) OTHER INFORMATION: /mod_base= gm

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 19
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "2'-O-methyladenosine"

     (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 20
          (D) OTHER INFORMATION: /mod_base= cm

     (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..20
          (D) OTHER INFORMATION: /note= "Target 2'-O-methyl
               oligonucleotide sequence"

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

NNNNNNNNNN NNNNNNNNNN                                                  20


(2) INFORMATION FOR SEQ ID NO:10:

      (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 8 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

     (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..8
          (D) OTHER INFORMATION: /note= "Matching DNA oligonucleotide
               probe"
```

-continued

```
    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CTTGCCAT                                                                   8

(2) INFORMATION FOR SEQ ID NO:11:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 8 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: other nucleic acid
          (A) DESCRIPTION: /desc = "2'-O-methyl oligonucleotide"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 1
          (D) OTHER INFORMATION: /mod_base= cm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 2
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 3
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 4
          (D) OTHER INFORMATION: /mod_base= gm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 5
          (D) OTHER INFORMATION: /mod_base= cm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 6
          (D) OTHER INFORMATION: /mod_base= cm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 7
          (D) OTHER INFORMATION: /mod_base= OTHER
                /note= "2'-O-methyladenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 8
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..8
          (D) OTHER INFORMATION: /note= "Matching 2'-O-methyl
                oligonucleotide analogue probe"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

NNNNNNNN                                                                   8

(2) INFORMATION FOR SEQ ID NO:12:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 8 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..8
          (D) OTHER INFORMATION: /note= "DNA oligonucleotide probe
               with 1 base mismatch"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CTTGCTAT                                                            8


(2) INFORMATION FOR SEQ ID NO:13:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 8 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: other nucleic acid
          (A) DESCRIPTION: /desc = "2'-O-methyl oligonucleotide"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 1
          (D) OTHER INFORMATION: /mod_base= cm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 2
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 3
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 4
          (D) OTHER INFORMATION: /mod_base= gm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 5
          (D) OTHER INFORMATION: /mod_base= cm

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 6
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 7
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "2'-O-methyladenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 8
          (D) OTHER INFORMATION: /mod_base= um

    (ix) FEATURE:
          (A) NAME/KEY: -
          (B) LOCATION: 1..8
          (D) OTHER INFORMATION: /note= "2'-O-methyl oligonucleotide
               analogue probe with 1 base mismatch"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

NNNNNNNN                                                            8


(2) INFORMATION FOR SEQ ID NO:14:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
```

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 10
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = cytosine covalently
modified at the 3' phosphate group with
a hexaethyleneglycol (HEG) linker"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

AAGATGNTAN 10

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 10
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = cytosine covalently modified
at the 3' phosphate group with a
hexaethyleneglycol (HEG) linker"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

AAGATNCTAN 10

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 10
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = cytosine covalently modified
at the 3' phosphate group with a
hexaethyleneglycol (HEG) linker"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

AAGANGCTAN 10

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 10
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = cytosine covalently modified

-continued

at the 3' phosphate group with a
hexaethyleneglycol (HEG) linker"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

AAGNTGCTAN 10

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 1
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = cytosine covalently modified
at the 5' phosphate group with a
fluorescein molecule"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

NTGAACGGTA GCATCTTGAC 20

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 11
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = adenine covalently modified
at the 3' phosphate group with a
hexaethyleneglycol (HEG) linker"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

AAAAANAAAA N 11

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 1
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = thymine covalently modified
at the 5' phosphate group with a
fluorescein molecule"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

NTTTTGTTTT T 11

-continued

```
(2) INFORMATION FOR SEQ ID NO:21:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 10
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = cytosine covalently modified
               at the 3' phosphate group with a
               hexaethyleneglycol (HEG) linker"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

CGCGCCGCGN                                                                 10


(2) INFORMATION FOR SEQ ID NO:22:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: other nucleic acid
          (A) DESCRIPTION: /desc = "2'-deoxynucleoside/nucleoside
               analogue decanucleotide probe"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 1
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 3
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 5
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 6
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 8
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 10
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine covalently
               modified at the 3' phosphate group with
               a hexaethyleneglycol (HEG) linker"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

NTNTNNTNTN                                                                 10
```

-continued

```
(2) INFORMATION FOR SEQ ID NO:23:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: other nucleic acid
          (A) DESCRIPTION: /desc = "2'-deoxynucleoside/nucleoside
               analogue decanucleotide probe"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 1
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 2
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 3
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 4
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 5
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 6
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 7
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 8
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 9
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 10
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2'-deoxyadenosine covalently
               modified at the 3' phosphate group with
               a hexaethyleneglycol (HEG) linker"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

NNNNNNNNNN                                                         10
```

```
(2) INFORMATION FOR SEQ ID NO:24:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: other nucleic acid
          (A) DESCRIPTION: /desc = "2'-deoxynucleoside/nucleoside
               analogue decanucleotide probe"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 1
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 3
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 5
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 6
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 8
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 10
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine
               covalently modified at the 3'
               phosphate group with a
               hexaethyleneglycol (HEG) linker"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

NTNTNNTNTN                                                          10


(2) INFORMATION FOR SEQ ID NO:25:

     (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 10 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: other nucleic acid
          (A) DESCRIPTION: /desc = "2'-deoxynucleoside/nucleoside
               analogue decanucleotide probe"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 1
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
          (A) NAME/KEY: modified_base
          (B) LOCATION: 2
          (D) OTHER INFORMATION: /mod_base= OTHER
               /note= "N = 5-propynyl-2'-deoxyuridine"
    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 3
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 4
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 5
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 6
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 7
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 8
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 2-amino-2'-deoxyadenosine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 9
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 5-propynyl-2'-deoxyuridine"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 10
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = 2-amino-2'-deoxyadenosine
              covalently modified at the 3'
              phosphate group with a
              hexaethyleneglycol (HEG) linker"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

NNNNNNNNNN                                                          10


(2) INFORMATION FOR SEQ ID NO:26:

    (i) SEQUENCE CHARACTERISTICS:
         (A) LENGTH: 10 base pairs
         (B) TYPE: nucleic acid
         (C) STRANDEDNESS: single
         (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 1
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = thymine covalently modified
              at the 5' hydroxyl group with a
              fluorescein molecule"

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 10
         (D) OTHER INFORMATION: /mod_base= OTHER
```

-continued

/note= "N = thymine covalently modified
at the 3' phosphate group with a
hexaethyleneglycol (HEG) linker which is
covalently bound to the 5' phosphate
group of the 5' guanine (N in pos. 1) of
SEQ ID NO:27"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

NATATTATAN 10

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 1
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = guanine covalently modified
at the 5' phosphate group with a
hexaethyleneglycol (HEG) linker which is
covalently bound to the 3' phosphate
group of the 3' thymine (N in pos. 10)
of SEQ ID NO:26"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

NCGCGGCGCG 10

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 15 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 6..10
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = guanine (G),
2',3'-dideoxyguanine (ddG),
2'-deoxyinosine (dI) or thymine (T)"

(ix) FEATURE:
(A) NAME/KEY: modified_base
(B) LOCATION: 15
(D) OTHER INFORMATION: /mod_base= OTHER
/note= "N = cytosine covalently modified
at the 5' phosphate group with a
hexaethyleneglycol (HEG) linker"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

TGGGCNNNNN TTGTN 15

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

-continued

```
ks(ii) MOLECULE TYPE: DNA

    (ix) FEATURE:
         (A) NAME/KEY: modified_base
         (B) LOCATION: 1
         (D) OTHER INFORMATION: /mod_base= OTHER
              /note= "N = cytosine covalently modified
              at the 5' phosphate group with a
              fluorescein molecule"

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

NAATACAACC CCCGCCCATC C                                   21
```

What is claimed is:

1. A composition for analyzing interactions between oligonucleotide targets and oligonucleotide probes comprising an array of a plurality of oligonucleotide analogue probes having different sequences, wherein said oligonucleotide analogue probes are coupled to a solid substrate at known locations and wherein said plurality of oligonucleotide analogue probes are selected to bind to complementary oligonucleotide targets with a similar hybridization stability across the array.

2. The composition of claim 1, wherein at least one of said oligonucleotide analogue probes is selected to maintain hybridization specificity or mismatch discrimination with said complementary oligonucleotide targets.

3. The composition of claim 1, wherein at least one of said oligonucleotide analogue probes has increased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

4. The composition of claim 1, wherein at least one of said oligonucleotide analogue probes has decreased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

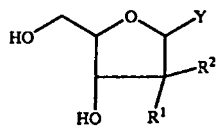
5. The composition of claim 2, wherein at least one of said oligonucleotide analogue probes has increased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

6. The composition of claim 2, wherein at least one of said oligonucleotide analogue probes has decreased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

7. The composition of claims 1–5 or 6, wherein said solid substrate is selected from the group consisting of silica, polymeric materials, glass, beads, chips, and slides.

8. The composition of claims 1–5 or 6, wherein said composition comprises an array of oligonucleotide analogue probes 5 to 20 nucleotides in length.

9. The composition of claims 1–5 or 6, wherein said array of oligonucleotide analogue probes comprises a nucleoside analogue with the formula



wherein:

the nucleoside analogue is not a naturally occurring DNA or RNA nucleoside;

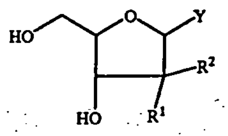
R^1 is selected from the group consisting of hydrogen, methyl, hydroxyl, alkoxy, alkythio, halogen, cyano, and azido;

R^2 is selected from the group consisting of hydrogen, methyl, hydroxyl, alkoxy, alkythio, halogen, cyano, and azido;

Y is a heterocyclic moiety;

and wherein said nucleoside analogue is incorporated into the oligonucleotide analogue by attachment to a 3' hydroxyl of the nucleoside analogue, to a 5' hydroxyl of the nucleoside analogue, or both the 3' nucleoside and the 5' hydroxyl of the nucleoside analogue.

10. The composition of claims 1–5 or 6, wherein said array of oligonucleotide analogue probes comprises a nucleoside analogue with the formula



wherein:

the nucleoside analogue is not a naturally occurring DNA or RNA nucleoside;

R^1 is selected from the group consisting of hydrogen, hydroxyl, methyl, methoxy, ethoxy, propoxy, allyloxy, propargyloxy, Fluorine, Chlorine, and Bromine;

R^2 is selected from the group consisting of hydrogen, hydroxyl, methyl, methoxy, ethoxy, propoxy, allyloxy, propargyloxy, Fluorine, Chlorine, and Bromine; and

Y is a base selected from the group consisting of purines, purine analogues pyrimidines, pyrimidine analogues, 3-nitropyrrole and 5-nitroindole;

and wherein said nucleoside analogue is incorporated into the oligonucleotide analogue by attachment to a 3' hydroxyl of the nucleoside analogue, to a 5' hydroxyl of the nucleoside analogue, or both the 3' nucleoside and the 5' hydroxyl of the nucleoside analogue.

11. The composition of claims 1–5 or 6, wherein each probe of said plurality of oligonucleotide analogue probes has at least one oligonucleotide analogue, and wherein at least one of said oligonucleotide analogues comprises a peptide nucleic acid.

12. The composition of claims 1–5 or 6, wherein at least one of said plurality of oligonucleotide analogue probes said array of oligonucleotide analogue probes is resistant to RNAase A.

13. The composition of claims 1–5 or 6, wherein said solid substrate is attached to over 1000 different oligonucleotide analogue probes.

14. The composition of claims 1–5 or 6, wherein each probe of said plurality of oligonucleotide analogue probes has at least one oligonucleotide analogue, and wherein at least one of said oligonucleotide analogues comprises 2'-O-methyl nucleotides.

15. The composition of claims 1–5 or 6, wherein said array of oligonucleotide analogue probes and said solid substrate comprises a plurality of different oligonucleotide analogue probes, each oligonucleotide analogue probes having the formula:

$$Y—L^1—X^1—L^2—X^2$$

wherein,

Y is a solid substrate;

X^1 and X^2 are complementary oligonuctèotides containing at least one nucleotide analogue;

L^1 is a spacer;

L^2 is a linking group having sufficient length such that X^1 and X^2 form a double-stranded oligonucleotide.

16. The composition of claim 15, wherein said composition comprises a library of unimolecular double-stranded oligonucleotide analogue probes.

17. The composition of claims 1–5 or 6, wherein said array of oligonucleotide analogue probes comprises a conformationally restricted array of oligonucleotide analogue probes with the formula:

$$—X^{11}—Z—X^{12}$$

wherein X^{11} and X^{12} are complementary oligonucleotides or oligonucleotide analogues and Z is a presented moiety.

18. The composition of claims 1–5 or 6, wherein each probe of said plurality of oligonucleotide analogue probes has at least one oligonucleotide analogue, and wherein at least one of said oligonucleotide analogues comprises a nucleotide with a base selected from the group of bases consisting of 5-propynyluracil, 5-propynylcytosine, 2-aminoadenine, 7-deazaguanine, 2-aminopurine, 8-aza-7-deazaguanine, 1H-purine, and hypoxanthine.

19. The composition of claims 1–5 or 6, wherein said plurality of oligonucleotide analogue probes are coupled to said solid substrate by light-directed chemical coupling.

20. The composition of claim 19, wherein said solid substrate is derivitized with a silane reagent prior to synthesis of said plurality of oligonucleotide analogue probes.

21. The composition of claims 1–5 or 6, wherein said plurality of oligonucleotide analogue probes are coupled to said solid substrate by flowing oligonucleotide analogue reagents over known locations of the solid substrate.

22. The composition of claim 21, wherein said solid substrate is derivitized with a silane reagent prior to synthesis of said plurality of oligonucleotide analogue probes.

23. The composition of claims 1–5 or 6, wherein at least one of plurality of said oligonucleotide analogue probes forms a first duplex with a target oligonucleotide sequence, wherein said oligonucleotide analogue probe has a corresponding oligonucleotide sequence that forms a second duplex with said target oligonucleotide sequence, wherein said second duplex is rich in A-T or G-C nucleotide pairs, and wherein said oligonucleotide analogue probe has at least one nucleotide analogue in place of an A, T, G, or C nucleotide of said corresponding oligonucleotide sequence at a position within said oligonucleotide analogue probe such that said first duplex has an increased hybridization stability than said second duplex.

24. The composition of claim 23, wherein said oligonucleotide analogue probe contains fewer bases than said corresponding oligonucleotide sequence.

25. The composition of claims 1–5 or 6, wherein said oligonucleotide analogue probe forms a first duplex with a target oligonucleotide sequence, wherein said oligonucleotide analogue probe has a corresponding oligonucleotide sequence that forms a second duplex with said target polynucleotide sequence, and wherein said oligonucleotide analogue probe is shorter than said corresponding polynucleotide sequence.

26. A composition for analyzing the interaction between an oligonucleotide target and an oligonucleotide probe comprising an array of a plurality of oligonucleotide probes having different sequences hybridized to complementary oligonucleotide analogue targets, wherein said oligonucleotide analogue targets bind to complementary oligonucleotide probes with a similar hybridization stability across the array.

27. The composition of claim 26, wherein at least one of said oligonucleotide analogue target is selected to maintain hybridization specificity or mismatch discrimination with said complementary oligonucleotide probes.

28. The composition of claim 26, wherein at least one of said oligonucleotide analogue targets has increased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

29. The composition of claim 26, wherein at least one of said oligonucleotide analogue targets has decreased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

30. The composition of claim 27, wherein at least one of said oligonucleotide analogue targets has increased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

31. The composition of claim 27, wherein at least one of said oligonucleotide analogue targets has decreased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

32. The composition of claims 26–30 or 31, wherein the oligonucleotide analogue target is a PCR amplicon.

33. The composition of claims 26–30 or 31, wherein at least one of said plurality of oligonucleotide probes comprise at least one oligonucleotide analogue.

34. The composition of claims 26–30 or 31, wherein at least one target oligonucleotide analogue acid is an RNA nucleic acid.

35. A method analyzing interactions between an oligonucleotide target and an oligonucleotide probe, comprising the steps of:

(a). synthesizing an oligonucleotide analogue array comprising a plurality of oligonucleotide analogue probes having different sequences, wherein said oligonucleotide analogue probes are coupled to a solid substrate at known locations, said solid substrate having a surface;

(b). exposing said oligonucleotide analogue probe array to a plurality of oligonucleotide targets under hybridization conditions such that said plurality of oligonucleotide analogue probes bind to complementary oligonucleotide targets with a similar hybridization stability across the array; and

(c). determining whether an oligonucleotide analogue probe of said oligonucleotide analogue probe array binds to at least one of said target nucleic acids.

36. The method in accordance of claim 35, wherein at least one of said oligonucleotide analogue probes is selected to maintain hybridization specificity or mismatch discrimination with said complementary oligonucleotide targets.

37. The method in accordance of claim 35, wherein at least one of said oligonucleotide analogue probes has increased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

38. The method in accordance of claim 35, wherein at least one of said oligonucleotide analogue probes has decreased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

39. The method in accordance of claim 36, wherein at least one of said oligonucleotide analogue probes has increased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

40. The method in accordance of claim 36, wherein at least one of said oligonucleotide analogue probes has decreased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

41. The method of claims 35–39 or 40, wherein said oligonucleotide target is selected from the group comprising genomic DNA, cDNA, unspliced RNA, mRNA, and rRNA.

42. The method of claims 35–39 or 40, wherein said target nucleic acid is amplified prior to said hybridization step.

43. The method of claims 35–39 or 40, wherein said plurality of oligonucleotide analogue probes is synthesized on said solid support by light-directed synthesis.

44. The method of claims 35–39 or 40, wherein said plurality of said oligonucleotide analogue probes is synthesized on said solid support by causing oligonucleotide analogue synthetic reagents to flow over known locations of said solid support.

45. The method of claims 35–39 or 40, wherein said step (a). comprises the steps of:

i). forming a plurality of channels adjacent to the surface of said substrate;

ii). placing selected reagents in said channels to synthesize oligonucleotide analogue probes at known locations; and

iii). repeating steps i). and ii). thereby forming an array of oligonucleotide analogue probes having different sequences at known locations on said substrate.

46. The method of claims 35–39 or 40, wherein said solid substrate is selected from the group consisting of beads, slides, and chips.

47. The method of claims 35–39 or 40, wherein said solid substrate is comprised of materials selected from the group consisting of silica, polymers and glass.

48. The method of claims 35–39 or 40, wherein the oligonucleotide analogue probes of said array are synthesized using photoremovable protecting groups.

49. The method of claims 35–39 or 40, further comprising selectively incorporating MeNPoc onto the 3' or 5' hydroxyl of at least one nucleoside analogue and selectively incorporating said nucleoside analogue into at least one of said oligonucleotide analogue probes.

50. The method of claims 35–39 or 40, wherein at least one of said oligonucleotide analogue probes is synthesized from phosphoramidite nucleoside reagents.

51. A method of detecting an oligonucleotide target, comprising enzymatically copying an oligonucleotide target using at least one nucleotide analogue, thereby producing multiple oligonucleotide analogue targets, selecting said oligonucleotide analogue targets such that said oligonucleotide analogue targets bind to the complementary oligonucleotide probes coupled to a solid surface at known locations of an array with a similar hybridization stability across the array, hybridizing the oligonucleotide analogue targets to complementary oligonucleotide probes, and detecting whether at least one of said oligonuclotide analogue targets binds to said complementary oligonucleotide acid probe.

52. The method of claim 51, wherein at least one of said oligonucleotide analogue targets is selected to maintain hybridization specificity or mismatch discrimination with said complementary oligonucleotide probes.

53. The method of claim 51, wherein at least one of said oligonucleotide analogue targets has increased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

54. The method of claim 51, wherein at least one of said oligonucleotide analogue targets has decreased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

55. The method of claim 52, wherein at least one of said oligonucleotide analogue targets has increased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

56. The method of claim 52, wherein at least one of said oligonucleotide analogue targets has decreased the thermal stability between said oligonucleotide analogue target and said complementary oligonucleotide probe as compared to an oligonucleotide target that is the perfect complement to the complementary oligonucleotide probe with which said oligonucleotide analogue target anneals.

57. The method of claims 51–55 or 56, wherein the oligonucleotide probe array comprises at least one oligonucleotide analogue probe which is complementary to at least one of said oligonucleotide analogue targets.

58. A method of making an array of oligonucleotide probes, comprising providing a plurality of oligonucleotide analogue probes having at least one oligonucleotide analogue, said oligonucleotide analogue probes having different sequences at known locations on an array, selecting the oligonucleotide analogue probes to hybridize with complementary oligonucleotide target sequences under hybridization conditions such that said oligonucleotide analogue probes bind to complementary oligonucleotide targets with a similar hybridization stability, across the array.

59. The method of claim 58, wherein at least one of said oligonucleotide analogue probes is selected to maintain hybridization specificity or mismatch discrimination with said complementary oligonucleotide targets.

60. The method of claim 58, wherein at least one of said oligonucleotide analogue probes has increased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

61. The method of claim 58, wherein at least one of said oligonucleotide analogue probes has decreased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

62. The method of claim 59, wherein at least one of said oligonucleotide analogue probes has increased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

63. The method of claim 59, wherein at least one of said oligonucleotide analogue probes has decreased the thermal stability between said oligonucleotide analogue probe and said complementary oligonucleotide target as compared to an oligonucleotide probe that is the perfect complement to the complementary oligonucleotide target with which said oligonucleotide analogue probe anneals.

64. The method in accordance with claims 58–62, or 63, further comprising incorporating at least one oligonucleotide analogue into at least one of the oligonucleotide analogue probes of the array to reduce or prevent the formation of secondary structure in the oligonucleotide of the array.

65. The method in accordance with claims 58–62, or 63, further comprising incorporating at least one oligonucleotide analogue into at least one of the oligonucleotide target to reduce or prevent the formation of secondary structure in the target polynucleotide sequence.

66. The method in accordance with claims 58–62, or 63, further comprising incorporating at least one oligonucleotide analogue into at least one of the oligonucleotide analogue probes of the array to create secondary structure in the oligonucleotide of the array.

67. The method in accordance with claims 58–62, or 63, further comprising incorporating a base selected from the group consisting of 5-propynyluracil, 5-propynylcytosine, 2-aminoadenine, 7-deazaguanine, 2-aminopurine, 8-aza-7-deazaguanine, 1H-purine, and hypoxanthine into the oligonucleotide analogue probes of the array.

68. The method of claim 67 further comprising selecting said at least one oligonucleotide analogue such that the oligonucleotide analogue probe is a homopolymer.

69. The method in accordance with claims 58–62, or 63, further comprising selecting said at least one oligonucleotide analogue from the group consisting essentially of oligonucleotide analogues comprising 2'-O-methyl nucleotides and oligonucleotides comprising a base selected from the group of bases consisting of 5 -propynyluracil, 5-propynylcytosine, 7-deazaguanine, 2-aminoadenine, 8-aza-7-deazaguanine, 1H-purine, and hypoxanthine.

70. The method in accordance with claims 58–62 or 63, further comprising selecting said at least one oligonucleotide analogue such that oligonucleotide analogue probes comprises at least one peptide nucleic acid.

71. The method in accordance with claims 58–62, or 63, further comprising selecting said at least one oligonucleotide analogue to increase image brightness when the oligonucleotide target and the oligonucleotide analogue probe hybridize in the presence of a fluorescent indicator, in comparison to a oligonucleotide probe without oligonucleotide analogs.

72. The method in accordance with claims 58–62, or 63, further comprising providing said plurality of oligonucleotide analogue probes in an array with at least 1000 other oligonucleotide analogue probes.

* * * * *

BB

US006261776B1

(12) United States Patent

Pirrung et al.

(10) Patent No.: **US 6,261,776 B1**

(45) Date of Patent: ***Jul. 17, 2001**

(54) NUCLEIC ACID ARRAYS

(75) Inventors: Michael C. Pirrung, Durham, NC (US); J. Leighton Read; Stephen P. A. Fodor, both of Palo Alto, CA (US); Lubert Stryer, Stanford, CA (US)

(73) Assignee: Affymetrix, Inc., Santa Clara, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: 09/292,455

(22) Filed: Apr. 15, 1999

Related U.S. Application Data

(63) Continuation of application No. 09/129,470, filed on Aug. 4, 1998, which is a continuation of application No. 08/456,598, filed on Jun. 1, 1995, which is a division of application No. 07/954,646, filed on Sep. 30, 1992, now Pat. No. 5,445,934, which is a division of application No. 07/850,356, filed on Mar. 12, 1992, now Pat. No. 5,405,783, which is a division of application No. 07/492,462, filed on Mar. 7, 1990, now Pat. No. 5,143,854, which is a continuation-in-part of application No. 07/362,901, filed on Jun. 7, 1989, now abandoned.

(51) **Int. Cl.**^{7} **C12Q 1/68**; G01N 33/543; A61K 38/00; C07H 21/04; C07H 21/00

(52) **U.S. Cl.** **435/6**; 435/7.92; 435/7.94; 435/795; 435/969; 435/973; 436/518; 436/527; 436/807; 436/809; 530/334; 536/24.3; 536/24.32; 536/25.32

(58) **Field of Search** 435/6, 7.92, 7.94, 435/7.95, 969, 973; 436/518, 527, 807, 809; 530/334; 536/24.3, 25.3, 25.32

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,849,137	11/1974	Barzynski et al.	96/67
3,862,056	1/1975	Hartman	252/511
3,939,350	2/1976	Arwin et al.	250/365
4,072,576	2/1978	Arwin et al.	195/103.5 R
4,180,739	12/1979	Abu-Shumays	250/461
4,238,757	12/1980	Schenck	357/25
4,269,933	5/1981	Pazos	430/291
4,314,821	2/1982	Rice	23/230 B
4,327,073	4/1982	Huang	424/1
4,339,528	7/1982	Goldman	430/323
4,342,905	8/1982	Fujii et al.	250/201
4,373,071	2/1983	Itakura	525/375
4,405,771	9/1983	Jagur	528/266
4,444,878	4/1984	Paulus	435/7

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

2242394	3/1974	(DE).
3440141	5/1986	(DE).
3505287	3/1988	(DE).

(List continued on next page.)

OTHER PUBLICATIONS

Bannwarth "Gene Technology: A Challenge for a Chemist" Chimia, 41:302–317 (Sep. 1987).

(List continued on next page.)

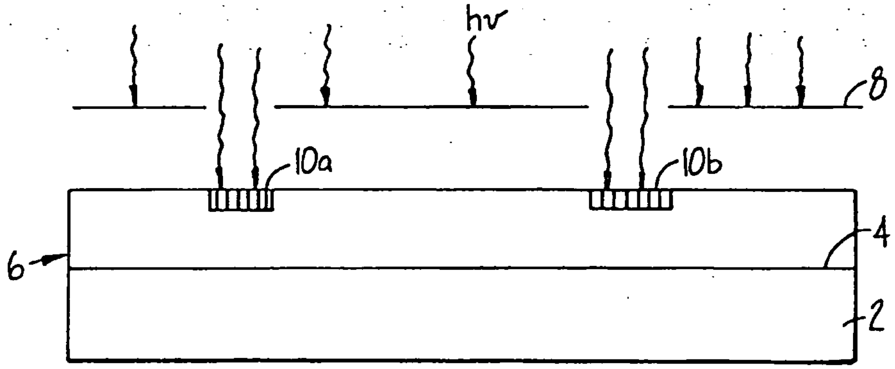
Primary Examiner—Jezia Riley

(74) *Attorney, Agent, or Firm*—Townsend and Townsend and Crew LLP

(57) **ABSTRACT**

A method and apparatus for preparation of a substrate containing a plurality of sequences. Photoremovable groups are attached to a surface of a substrate. Selected regions of the substrate are exposed to light so as to activate the selected areas. A monomer, also containing a photoremovable group, is provided to the substrate to bind at the selected areas. The process is repeated using a variety of monomers such as amino acids until sequences of a desired length are obtained. Detection methods and apparatus are also disclosed.

39 Claims, 20 Drawing Sheets



U.S. PATENT DOCUMENTS

4,444,892	4/1984	Malmros	436/528
4,448,534	5/1984	Wertz et al.	356/435
4,458,066	7/1984	Caruthers et al.	536/27
4,483,920	11/1984	Gillespie et al.	435/6
4,500,707	2/1985	Caruthers et al.	536/7
4,516,833	5/1985	Fusek	350/162.12
4,517,338	5/1985	Urdea et al.	525/54.11
4,537,861	8/1985	Elings et al.	436/518
4,542,102	9/1985	Dattagupta et al.	435/6
4,555,490	11/1985	Merril	436/86
4,562,157	12/1985	Lowe et al.	435/291
4,569,967	2/1986	Kornreich et al.	525/54.11
4,580,895	4/1986	Patel	356/39
4,584,277	4/1986	Ullman	436/509
4,613,566	9/1986	Potter	435/6
4,624,915	11/1986	Schindler et al.	435/4
4,626,684	12/1986	Landa	250/328
4,631,211	12/1986	Houghten	428/35
4,637,861	1/1987	Krull et al.	204/1
4,677,054	6/1987	White et al.	435/6
4,681,859	7/1987	Kramer	436/501
4,683,202	7/1987	Mullis	435/91
4,689,405	8/1987	Frank et al.	536/27
4,704,353	11/1987	Humphries et al.	435/4
4,711,955	12/1987	Ward et al.	536/29
4,713,326	12/1987	Dattagupta et al.	435/6
4,713,347	12/1987	Mitchell et al.	436/501
4,719,615	1/1988	Feyrer et al.	369/284
4,722,906	2/1988	Guire	436/501
4,728,502	3/1988	Hamill	422/116
4,728,591	3/1988	Clark et al.	430/5
4,731,325	3/1988	Palva et al.	435/5
4,755,458	7/1988	Rabbani et al.	435/5
4,762,881	8/1988	Kauer	525/54.11
4,777,019	10/1988	Dandekar	422/68
4,780,504	10/1988	Buendia et al.	525/54.11
4,786,170	11/1988	Groebler	356/318
4,786,684	11/1988	Glass	525/54.1
4,794,150	12/1988	Steel	525/54.11
4,808,508	2/1989	Platzer	430/143
4,810,869	3/1989	Yabe et al.	250/501
4,811,062	3/1989	Tabata et al.	356/152
4,812,512	3/1989	Buendia et al.	525/54.11
4,820,630	4/1989	Taub	435/5
4,822,566	4/1989	Newman	422/68
4,833,092	5/1989	Geysen	436/501
4,844,617	7/1989	Kelderman et al.	356/372
4,846,552	7/1989	Veldkamp et al.	350/162.2
4,849,513	7/1989	Smith et al.	536/27
4,855,225	8/1989	Fung et al.	435/6
4,865,990	9/1989	Stead et al.	435/803
4,868,103	9/1989	Stavrianopoulos et al.	435/5
4,874,500	10/1989	Madou et al.	204/412
4,886,741	12/1989	Schwartz	435/5
4,888,278	12/1989	Singer et al.	435/6
4,923,901	5/1990	Koester et al.	521/53
4,925,785	5/1990	Wang et al.	435/6
4,946,942	8/1990	Fuller et al.	530/335
4,973,493	11/1990	Guire	427/2
4,979,959	12/1990	Guire	623/66
4,981,783	1/1991	Augenlicht	435/6
4,981,985	1/1991	Kaplan et al.	556/50
4,984,100	1/1991	Takayama et al.	360/49
4,987,065	1/1991	Stavrianopoulos et al.	435/5
4,988,617	1/1991	Landegren et al.	435/6
4,992,583	2/1991	Farnsworth	436/89
4,994,373	2/1991	Stavrianopoulos et al.	435/6
5,002,867	3/1991	Macevicz	435/6
5,021,550	6/1991	Zeiger	530/334
5,026,773	6/1991	Steel	525/54.11

5,026,840	6/1991	Dattagupta et al.	536/27
5,028,525	7/1991	Gray et al.	435/6
5,043,265	8/1991	Tanke et al.	435/6
5,047,524	9/1991	Andrus et al.	536/27
5,079,600	1/1992	Schnur et al.	357/4
5,081,584	1/1992	Omichinski et al.	364/497
5,082,830	1/1992	Brakel et al.	514/44
5,091,652	2/1992	Mathies et al.	250/458.1
5,112,962	5/1992	Letsinger et al.	536/27
5,141,813	8/1992	Nelson	428/402
5,143,854	9/1992	Pirrung et al.	436/518
5,153,319	10/1992	Caruthers et al.	536/27
5,192,980	3/1993	Dixon et al.	356/326
5,200,051	4/1993	Cozzette et al.	204/403
5,202,231	4/1993	Drmanac et al.	435/6
5,206,137	4/1993	Ip et al.	435/6
5,215,882	6/1993	Bahl et al.	435/6
5,215,889	6/1993	Schultz	435/41
5,232,829	8/1993	Longiaru et al.	435/6
5,235,028	8/1993	Barany et al.	528/335
5,242,974	9/1993	Holmes	525/54.11
5,252,743	10/1993	Barrett et al.	548/303.7
5,256,549	10/1993	Urdea et al.	435/91
5,258,506	11/1993	Urdea et al.	536/23.1
5,306,641	4/1994	Saccocio	436/85
5,310,893	5/1994	Erlich et al.	536/24.31
5,324,633	6/1994	Fodor et al.	435/6
5,348,855	9/1994	Dattagupta et al.	435/6
5,384,261	1/1995	Winkler et al.	436/518
5,405,783	4/1995	Pirrung et al.	436/518
5,424,186	6/1995	Fodor et al.	435/6
5,436,327	7/1995	Southern et al.	536/25.34
5,445,934	8/1995	Fodor et al.	435/6
5,447,841	9/1995	Gray et al.	435/6
5,486,452	1/1996	Gordon et al.	435/5
5,489,507	2/1996	Chehab	435/6
5,489,678	2/1996	Fodor et al.	536/22.1
5,492,806	2/1996	Drmanac et al.	435/5
5,510,270	4/1996	Fodor et al.	436/518
5,525,464	6/1996	Dramanac et al.	435/6
5,527,681	6/1996	Holmes	435/6
5,552,270	9/1996	Khrapko et al.	435/6
5,556,961	9/1996	Foote et al.	536/27.1
5,561,071	10/1996	Hollenberg et al.	437/1
5,571,639	11/1996	Hubbell et al.	430/5
5,593,839	1/1997	Hubbell et al.	435/6
5,653,939	8/1997	Hollis et al.	422/50
5,667,667	9/1997	Southern	205/687
5,667,972	9/1997	Drmanac et al.	435/6
5,695,940	12/1997	Drmanac et al.	435/6
5,698,393	12/1997	Macioszek et al.	435/5
5,700,637	12/1997	Southern	435/6
5,707,806	1/1998	Shuber	435/6
5,744,305	4/1998	Fodor et al.	435/6
5,777,888	7/1998	Rine et al.	364/496
5,800,992	9/1998	Fodor et al.	435/6
5,807,522	9/1998	Brown et al.	422/50
5,830,645	11/1998	Pinkel et al.	435/287.1
5,843,767	12/1998	Beattie	435/6
5,846,708	12/1998	Hollis et al.	435/6
5,871,697	2/1999	Rothberg et al.	422/68.1
5,972,619	10/1999	Drmanac et al.	435/6
6,018,041	1/2000	Drmanac et al.	536/24.3
6,025,136	2/2000	Drmanac et al.	435/6
6,054,270	4/2000	Southern	435/6

FOREIGN PATENT DOCUMENTS

046 083	2/1982	(EP) .
088 636	9/1983	(EP) .
103 197	3/1984	(EP) .
127 0438	12/1984	(EP) .

063 810	3/1986	(EP) .
194 132	9/1986	(EP) .
228 075	7/1987	(EP) .
245 662	11/1987	(EP) .
268 237	5/1988	(EP) .
281 927	9/1988	(EP) .
228 310	10/1988	(EP) .
288 310	10/1988	(EP) .
304 202	2/1989	(EP) .
307 476	3/1989	(EP) .
319 012	6/1989	(EP) .
328 256	8/1989	(EP) .
333 561	9/1989	(EP) .
337 498	10/1989	(EP) .
386 229	4/1990	(EP) .
373 203	6/1990	(EP) .
392 546	10/1990	(EP) .
173 339	1/1992	(EP) .
171 150	3/1992	(EP) .
237 362	3/1992	(EP) .
185 547	6/1992	(EP) .
260 634	6/1992	(EP) .
232 967	4/1993	(EP) .
235 726	5/1993	(EP) .
476 014	8/1994	(EP) .
225 807	10/1994	(EP) .
717 113	6/1996	(EP) .
848 067	6/1998	(EP) .
619 321	1/1999	(EP) .
2156074	3/1988	(GB) .
2559783	3/1988	(FR) .
2196476	4/1988	(GB) .
2248840	9/1992	(GB) .
49-110601	10/1974	(JP) .
60-248669	12/1985	(JP) .
63-084499	4/1988	(JP) .
63-223557	9/1988	(JP) .
1-233447	9/1989	(JP) .
WO 84/03151	8/1984	(WO) .
WO 84/03564	9/1984	(WO) .
WO 85/01051	3/1985	(WO) .
WO 86/00991	2/1986	(WO) .
WO 86/06487	11/1986	(WO) .
8810400	5/1988	(WO) .
WO 88/04777	6/1988	(WO) .
WO 89/05616	6/1989	(WO) .
WO 89/08834	9/1989	(WO) .
WO 89/11548	11/1989	(WO) .
WO 89/10977	11/1989	(WO) .
WO 89/12819	12/1989	(WO) .
WO 90/00887	2/1990	(WO) .
WO 90/15070	2/1990	(WO) .
WO 90/03382	4/1990	(WO) .
WO 90/04652	5/1990	(WO) .
WO 91/04266	4/1991	(WO) .
WO 91/07087	5/1991	(WO) .
WO 92/16655	1/1992	(WO) .
WO 92/10092	6/1992	(WO) .
WO 92/10588	6/1992	(WO) .
WO 93/02992	2/1993	(WO) .
WO 93/09668	5/1993	(WO) .
WO 88/01302	6/1993	(WO) .
WO 93/11262	6/1993	(WO) .
WO 93/22480	11/1993	(WO) .
WO 93/22546	11/1993	(WO) .
WO 95/11995	5/1995	(WO) .
WO 95/33846	12/1995	(WO) .
WO 96/23078	8/1996	(WO) .
WO 97/10365	3/1997	(WO) .
WO 97/17317	5/1997	(WO) .
WO 97/19410	5/1997	(WO) .
WO 97/27317	7/1997	(WO) .
WO 97/29212	8/1997	(WO) .
WO 98/31836	7/1998	(WO) .

OTHER PUBLICATIONS

Bannwarth et al. "A System for the Simultaneous Chemical Synthesis of Different DNA Fragments on Solid Support" DNA, 5:413–419 (Oct. 1986).

Sequencing by Hybridization Workshop, listing of participants and workshop presentation summaries (1991).

"A Sequencing Reality Check," *Science* 242:1245 (1988).

"Affymax raises $25 million to develop high–speed drug discovery system," *Biotechnology News*, 10(3):7–8 (1990).

"Preparation of fluorescent–labeled DNA and its use as a probe in molecular hybridization," *Bioorg Khim*, 12(11):1508–1513 (1986).

Abbott et al., "Manipulation of the Wettability of Surfaces on the 0.1– to 1–Micrometer Scale Through Micromachining and Molecular Self–Assembly," *Science*, 257:1380–1382 (1992).

Adams et al., "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project," *Science*, 252(5013):1651–1656 (1991).

Adams et al., "Photolabile Chelators That "Cage" Calcium with Improved Speed of Release and Pre–Photolysis Affinity," *J. Gen. Physiol.*, p. 9a (Dec. 1986).

Adams et al., "Biologically Useful Chelators That Take Up Ca2+ upon Illumination," *J. Am. Chem. Soc.*, 111:7957–7968 (1989).

Amit et al., "Photosensitive Protecting Groups of Amino Sugars and Their Use in Glycoside Synthesis. 2–Nitrobenzyloxycarbonylamino and 6–Nitroveratryloxycarbonylamino Derivatives," *J. Org. Chem*, 39(2):192–196 (1974).

Amit et al., "Photosensitive Protecting Groups—A Review," *Israel J. Chem.*, 12(1–2):103–113 (1974).

Applied Biosystems, Model 431A Peptide Synthesizer User's manual, Sections 2 and 6, (Aug. 15, 1989).

Ajayaghosh et al., "Solid–Phase Synthesis of N–Methyl– and N–Ethylamides of Peptides Using Photolytically Detachable ((3–Nitro–4((alkylamino)methyl)benzamido)methyl)polystyrene Resin," *J.Org.Chem.*, 55(9):2826–2829 (1990).

Ajayaghosh et al., "Solid–phase synthesis of C–terminal peptide amides using a photoremovable α–methylphenacylamido anchoring linkage," *Proc. Ind. Natl. Sci* (*Chem.Sci.*), 100(5):389–396 (1988).

Ajayaghosh et al., "Polymer–supported Solid–phase Synthesis of C–Terminal Peptide N–Methylamides Using a Modified Photoremovable 3–Nitro–4–N–methylaminomethylpolystyrene Support," *Ind.J.Chem.*, 27B:1004–1008 (1988).

Ajayaghosh et al., "Polymer–Supported Synthesis of Protected Peptide Segments on a Photosensitive o–Nitro(α–Methyl)Bromobenzyl Resin," *Tetrahedron*, 44(21):6661–6666 (1988).

Arnold et al., "A Novel Universal Support for DNA & RNA Synthesis," abstract from Federation Proceedings, 43(7):abstract No. 3669 (1984).

Atherton et al., Solid Phase Peptide Synthesis: A Practical Approach, IRL Press, (1989), tbl of cont., pp. vii–ix.

Augenlicht et al., "Cloning and Screening of Sequences Expressed in a Mouse Colon Tumor," *Cancer Research*, 42:1088–1093 (1982).

Augenlicht et al., "Expression of Cloned Sequences in Biopsies of Human Colonic Tissue and in Colonic Carcinoma Cells Induced to Differentiate in Vitro," *Cancer Res.*, 47:6017–6021 (1987).

Bains, W., "Hybridization Methods for DNA Sequencing," *Genomics*, 11(2):294–301 (1991).

Bains et al., "A Novel Method for Nucleic Acid Sequence Determination," *J.Theor.Biol.*, 135:303–307 (1988).

Bains, W., "Alternative Routes Through the Genome," *Biotechnology*, 8:1251–1256 1990.

Balachander et al., "Functionalized Siloxy–Anchored Monolayers with Exposed Amino, Azido, Bromo, or Cyano Groups," *Tetrahed. Ltrs.*, 29(44):5593–5594 (1988).

Baldwin et al., "New Photolabile Phosphate Protecting Groups," *Tetrahed.*, 46(19):6879–6884 (1990).

Barltrop et al., "Photosensitive Protective Groups," *Chemical Communications*, pp. 822–823 (1966).

Barinaga, M., "Will 'DNA Chip' Speed Genome Initiative," *Science*, 253:1489 (1985).

Bart et al., "Microfabricated Electrohydrodynamic Pumps," *Sensors and Actuators*, A21–A23:193–197 (1990).

Bartsh et al., "Cloning of mRNA sequences from the human colon: Preliminary characterisation of defined mRNAs in normal and neoplastic tissues," *BrJ.Can.*, 54:791–798 (1986).

Baum, R., "Fledgling firm targets drug discovery process," *Chem. Eng. News*, pp. 10–11 (1990).

Beltz et al., "Isolation of Multigene Families and Determination of Homologies by Filter Hybridization Methods," *Methods in Enzymology*, 100:266–285 (1983).

Benschop, Chem. Abstracts 114(26):256643 (1991).

Bhatia et al., "New Approach To Producing Patterned Biomolecular Assemblies," *J. American Chemical Society*, 114:4432–4433 (1992).

Biorad Chromatography Electrophoresis Immunochemistry Molecular Biology HPLC catalog M 1987 p. 182.

Blawas et al., "Step–and–Repeat Photopatterning of Protein Features Using Caged–Biotin–BSA: Characterization and Resolution," *Langmuir*, 14(15):4243–4250 (1998).

Blawas, A.S., "Photopatterning of Protein Features using Caged–biotin–Bovine Serum Albumin," dissertation for Ph.D at Duke University in 1998.

Bos et al., "Amino–acid substirutions at coddon 13 of the N–ras oncogene in human acute myeloid leukaemia," *Nature*, 315:726–730 (1985).

Boyle et al., "Differential distribution of long and short interspersed element sequences in the mouse genome: Chromosome karyotyping by fluorescence in situ hybridization," *PNAS*, 87:7757–7761 (1990).

Brock et al., "Rapid fluorescence detection of in situ hybridization with biotinylated bovine herpesvirus–1 DNA probes," *J.Veterinary Diagnostic Invest.*, 1:34–38 (1989).

Burgi et al., "Optimization in Sample Stacking for High-Performance Capillary Electrophoresis," *Anal. Chem.*, 63:2042–2047 (1991).

Cameron et al., "Photogeneration of Organic Bases from o–Nitrobenzyl–Derived Carbamates," *J. Am. Chem. Soc.*, 113:4303–4313 (1991).

Carrano et al., "A High–Resolution, Fluorescence–Based, Semiautomated Method for DNA Fingerprinting," *Genomics*, 4:129–136 (1989).

Caruthers, M.H., "Gene Synthesis Machines: DNA Chemistry and Its Uses," *Science*, 230:281–285 (1985).

Chatterjee et al., "Inducible Alkylation of DNA Using an Oligonucleotide–Quinone Conjugate," *Am. J. Chem. Soc.*, 112:6397–6399 (1990).

Chee et al., "Accessing Genetic Information with High––Density DNA Arrays," *Science*, 274:610–614 (1996).

Chehab et al., "Detection of sicle cell anaemia mutation by colour DNA amplification," *Lancet*, 335:15–17 (1990).

Chehab et al., "Detection of specific DNA sequences by fluorescence amplification: A color complementation assay," *PNAS*, 86:9178–9182 (1989).

Clevite Corp., Piezolelectric Technology, Data for Engineers.

Corbett et al., "Reaction of Nitroso Aromatics with Glyoxylic Acid. A New Path to Hydroxamic Acids," *J. Org. Chem.*, 45:2834–2839 (1980).

Craig et al., "Ordering of cosmid clones covering the Herpes simplex virus type 1 (HSV–1) genome: a test case for fingerprinting by hybridization," *Nuc. Acid. Res.*, 18(9):2653–2660 (1990).

Cummings et al., "Photoactivable Fluorophores. I. Synthesis and Photoactivation of o–Nitrobenzyl–Quenched Fluorescent Carbamates," *Tetrahedron Letters*, 29(1):65–68 (1988).

Diggelmann, "Investigating the VLSIPS synthesis process," Sep. 9, 1994.

Di Mauro et al., "DNA Technology in Chip Construction," *Adv. Mater.*, 5(5):384–386 (1993).

Drmanac et al., "Partial Sequencing by Oligo–Hybridization Concept and Applications in Genome Analysis," 1st Int. Conf. Electrophor., Supercomp., Hum. Genome pp. 60–74 (1990).

Drmanac et al., "Sequencing by Oligonucleotide Hybridization: A Promising Framework in Decoding of the Genome Program?," 1st Int. Conf. Electrophor., Supercomp., Hum. Genome pp. 47–59 (1990).

Drmanac et al., "Laboratory Methods, Reliable Hybridization of Oligonucleotides as Short as Six Nucleotides," *DNA and Cell Biol.*, 9(7):527–534 (1990).

Drmanac et al., "Sequencing of Megabase Plus DNA by Hybridization: theory of the Method," *Genomics*, 4:114–128 (1989).

Dramanac et al., "Sequencing of Megabase Plus DNA by Hybridization: Theory of the Method," abstract of presentation given at Cold Spring Harbor Symposium on Genome Mapping and Sequencing, Apr. 27, 1988 thru May 1, 1988.

Dulcey et al., "Deep UV Photochemistry of Chemisorbed Monolayers: Patterned Coplanar Molecular Assemblies," *Science*, 252:551–554 (1991).

Duncan et al., "Affinity Chromatography of a Sequence–Specific DNA Binding Protein Using Teflon–Linked Oligonucleotides," *Analytical Biochemistry*, 169:104–108 (1988).

Effenhauser et al., "Glass Chips for High–speed Capillary Electrophoresis Separations with Submicrometer Plate Heights," Anal. Chem., 65:2637–2642 (1993).

Effenhauser et al., "High–Speed Separation of Antisense Oligonucleotides on a Micromachined Capillary Electrophoresis Device," *Anal. Chem.*, 66:2949–2953 (1994).

Ekins et al., "High Specific Activity Chemiluminescent and Fluorescent Markers: their Potential Application to High Sensitivity and 'Multi–analyte' Immunoassays," *J. Bioluminescence Chemiluminescence*, 4:59–78 (1989).

Ekins et al., "Development of Microspot Multi–Analyte Rationmetric Immunoassay Using dual Fluorescent–Labelled Antibodies," *Anal. Chemica Acta*, 227:73–96 (1989).

Ekins et al., "Multianalyte Microspot Immunoassay–Microanalytical 'Compact Disk' of the Future," *Clin. Chem.*, 37(11):1955–1967 (1991).

Ekins, R.P., "Multi-Analyte immunoassay*," *J. Pharmaceut. Biomedical Analysis*, 7(2):155–168 (1989).

Evans et al., "Microfabrication for Automation of Molecular processes in Human Genome Analysis,"0 *Clin. Chem.*, 41(11):1681 (1995).

Evans et al., "Physical mapping of complex genomes by cosmid multiplex analysis," *PNAS*, 86:5030–5034 (1989).

Ezaki et al., "Small-Scale DNA Preparation for Rapid Genetic Identification of *campylobacter* Species with Radioisotope," *Microbiol. Immunology*, 32(2):141–150 (1988).

Fan et al., "Mapping small DNA sequences by fluorescence in situ hybridization directly on banded metaphase chromosomes," PNAS, 87(16):6223–6227 (1990).

Fan et al., "Micromachining of Capillary Electrophoresis Injectors and Separators on Glass Chips and Evaluation of Flow at Capillary Intersections," Anal. Chem., 66:177–184 (1994).

Fettinger et al., "Stacked modules for micro flow systems in chemical analysis: concept and studies using an enlarged model," *Sensors and Actuators*, B17:19–25 (1993).

Flanders et al., "A new interferometric alignment technique," *App. Phys. Ltrs.*, 31(7):426–429 (1977).

Fodor et al., "Multiplexed biochemical assays with biological chips," *Nature*, 364:555–556 (1993).

Fodor et al., "Light-directed, Spatially Addressable Parallel Chemical Synthesis," *Science*, 251:767–773 (1991).

Forman et al., "Thermodynamics of Duplex Formation and Mismatch Discrimination on Photolithographically Synthesized Oligonucleotide Arrays," chapter 13 pp. 206–228 from *Molecular Modeling of Nucleic Acids*, ACS Symposium Series 682, Apr. 13–17, 1997, Leontis et al., eds.

Frank et al., "Simultaneous Multiple Peptide Synthesis Under Continuous flow Conditions on Cellulose Paper Discs as Segmental Solid Supports," *Tetrahedron*, 44(19):6013–6040 (1988).

Frank et al., "Automation of DNA Sequencing Reactions and Related Techniques: A Workstation for Micromanipulation of Liquids," *Bio/Technology*, 6:1211–1212 (1988).

Frank et al., "Simultaneous Synthesis and Biological Applications of DNA Fragments: An Efficient and Complete Methodology," *Methods in Enzymology*, 154:221–250 (1987).

Fuhr et al., "Travelling wave-driven microfabricated electrohydrodynamic pumps for liquids," *J. Micromech. Microeng.*, 4:217–226 (1994).

Fuller et al., "Urethane-Protected Amino Acid N-Carboxy Anhydrides and Their Use in Peptide Synthesis," *J. Amer. Chem. Soc.*, 112(20):7414–7416 (1990).

Furka et al., "General method for rapid synthesis of multicomponent peptide mixtures," *Int. J. Peptide Protein Res.*, 37:487–493 (1991).

Furka et al., "Cornucopia of Peptides by Synthesis," 14th Int. Congress of Biochem. abst.#FR:013, Jul. 10–15, 1988 Prague, Czechoslovakia.

Furka et al., "More Peptides by Less Labour," abst. 288, Int. Symp. Med. Chem., Budapest Hungary Aug. 15–19, 1988.

Gait, eds., pp. 1–115 from *Oligonucleotide Synthesis: A Practical Approach*, IRL Press, (1984).

Gazard et al., "Lithographic Technique Using Radiation-Induced Grafting of Acrylic Acid into Poly(Methyl Methacrylate) Films," *Polymer Engineering and Science*, 20(16):1069–1072 (1980).

Gergen et al., "Filter replicas and permanent collections of recombinant DNA plasmids," *NucAcids Res.*, 7(8):2115–2137 (1979).

Getzoff et al., "Mechanisms of Antibody Binding to a Protein," *Science*, 235:1191–1196 (1987).

Geysen et al., "Strategies for epitope analysis using peptide synthesis," *J. Immunol. Meth.*, 102:259–274 (1987).

Geysen et al., "Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid," *PNAS*, 81:3998–4002 (1984).

Geysen et al., "A synthetic strategy for epitope mapping" from Peptides:Chem. & Biol., Proc. of 10th Am. Peptide Symp., May 23–28, 1987, pp. 519–523, (1987).

Geysen, "Antigen-antibody interactions at the molecular level: adventures in peptide synthesis," *Immunol. Today*, 6(12):364–369 (1985).

Geysen et al., "Cognitive Features of Continuous Antigenic Determinants," from Synthetic Peptides: Approaches to Biological Probes, pp. 19–30, (1989).

Geysen et al., "Chemistry of Antibody Binding to a Protein," *Science*, 235:1184–1190 (1987).

Geysen et al., "The delineation of peptides able to mimic assembled epitopes," 1986 CIBA Symp., pp. 130–149.

Geysen et al., "Cognitive Features of Continuous Antigenic Determinants," *Mol. Recognit.*, 1(1):1–10 (1988).

Geysen et al., "A Prio Ri Delineation of a Peptide Which Mimics A Discontinuous Antigenic Determinant," *Mol. Immunol.*, 23(7):709–715 (1986).

Gilon et al., "Backbone Cyclization: A New Method for Conferring Conformational Constraint on Peptides," *Biopolymers*, 31(6):745–750 (1991).

Gineras et al., "Hybridization properties of immobilized nucleic acids," *Nuc. Acids Res.*, 15(13):5373–5390 (87).

Gummerlock et al., "RAS Enzyme-Linked Immunoblot Assay Discriminates p21 Species: A Technique to Dissect Gene Family Expression," *Anal. Biochem.*, 180:158–168 (1989).

Gurney et al., "Activation of a potassium current by rapid photochemically generated step increases of intracellular calcium in rat sympathetic neurons," *PNAS*, 84:3496–3500 (1987).

Haase et al., "Detection of Two Viral Genomes in Single Cells by Double-Label Hybridization in Situ and Color Microradioautography," *Science*, 227:189–192 (1985).

Hacia, et al., "Two color hybridization analysis using high density oligonucleotide arrays and energy transfer dyes," *Nuc. Acids Res.*, 26(16):3865–3866 (1998).

Hack, M.L., "Conics Formed to Make Fluid & Industrial Gas Micromachines," *Genetic Engineering News*, 15(18):1, 29 (1995).

Hagedorn et al., "Pumping of Water Solutions in Microfabricatedd Electrohydrodynamic Systems," from Micro Electro Mechanical Systems conference in Travemunde Germany (1992).

Hanahan et al., "Plasmid Screening at High Colony Density," *Meth. Enzymology*, 100:333–342 (1983).

Hanahan et al., "Plasmid screening at high colony density," *Gene*, 10:63–67 (1980).

Haridasan et al., "Peptide Synthesis using Photolytically Cleavable 2–Nitrobenzyloxycarbonyl Protecting Group," *Proc. Indian Natn. Sci. Adad.*, 53A(6):717–728 (1987).

Harrison et al., "Capillary Electrophoresis and Sample Injection Systems Integrated on a Planar Glass Chip," *Anal. Chem.*, 64:1926–1932 (1992).

Harrison et al., "Micromachining a Minaturized Capillary Electrophoresis–Based Chemical Analysis System on a Chip," Science, 261:895–897 (1993).

Harrison et al., "Towards minaturized electrophoresis and chemical analysis systems on silicon: an alternative to chemical sensors*," *Sensors and Actuators*, B10:107–116 (1993).

Harrison et al., "Rapid separation of fluorescein derivatives using a micromachined capillary electrophoresis system," *Analytica Chemica Acta*, 283:361–366 (1993).

Hellberg et al., "Minimum analogue peptide sets (MAPS) for quantitative structure–activity relationships," *Int. J. Peptide Protein Res.*, 37:414–424 (1991).

Hilser et al., "Protein and peptide mobility in capillary zone electrophoresis, A comparison of existing models and further analysis," *J. Chromatography*, 630:329–336 (1993).

Ho et al., "Highly Stable Biosensor Using an Artificial Enzyme," *Anal.Chem.*, 59:536–537 (1987).

Hochgeschwender et al., "Preferential expression of a defined T–cell receptor β–chain gene in hapten–specific cytotoxic T–cell clones," *Nature*, 322:376–378 (1986).

Hodgson, J., "Assays A La Photolithography," *Biotech.*, 9:419 (1991).

Hopman et al., "Bi–color detection of two target DNAs by non–radioactive in situ hybridization*," *Histochem.*, 85:1–4 (1986).

Iwamura et al., "1–Pyrenylmethyl Esters, Photolabile Protecting Groups for Carboxlic Acids," *Tetrahedron Ltrs.*, 28(6):679–682 (1987).

Iwamura et al., "1–(α–Diazobenzyl)pyrene: A Reagent for Photolabile and Fluorescent Protection of Carboxyl Groups of Amino Acids and Peptides," *Synlett*, pp. 35–36 (1991).

Jacobson et al., "Effects of Injection Schemes and Column Geometry on the Performance of Microchip Electrophoresis Devices," Anal. Chem., 66:1107–1113 (1994).

Jacobsen et al., "Open Channel Electrochromatography on a Microchip," Anal. chem., 66:2369–2373 (1994).

Jacobson et al., "Microchip Capillary Electrophoresis with an Integrated Postcolumn Reactor" Anal. Chem., 66:3472–3476 (1994).

Jacobson et al., "Precolumn Reactions with Electrophoretic Analysis Integrated on a Microchip," *Anal. Chem.*, 66:4127–4132 (1994).

Jacobson et al., "Microfabricated chemical measurement systems," *Nature Medicine*, 1(10):1093–1096 (1995).

Jacobsen et al., "Fused Quartz Substrates for Microchip Electrophoresis," *Anal. chem.*, 67:2059–2063 (1995).

Jacobson et al., "High–Speed Separtions on a Microchip," Anal. Chem., 66:1114–1118 (1994).

Jacobson et al., "Microchip electrophoresis with sample stacking," *Electrophoresis*, 16:481–486 (1995).

Jayakumari, "Peptide synthesis in a triphasic medium catalysed by papain immobilized on a crosslinked polystyrene support," *Indian J. Chemistry*, 29B:514–517 (1990).

Kaiser et al., "Peptide and Protein Synthesis by Segment Synthesis–Condensation," *Science*, 243:187–192 (1989).

Kaplan et al., "Photolabile chelators for the rapid photorelease of divalent cations," *PNAS*, 85:6571–6575 (1988).

Karube, "Micro–biosensors based on silicon fabrication technology," chapter 25 from Biosensors:Fundamentals and Applications, Turner et al., eds., Oxford Publ., 1987, pp. 471–480 (1987).

Kates et al., "A Novel, Convenient, Three–dimensional Orthogonal Strategy for Solid–Phase Synthesis of Cyclic Peptides 1–3," *Tetrahed. Letters*, 34(10):1549–1552 (1993).

Kerkof et al., "A Procedure for Making Simultaneous Determinations of the Relative Levels of Gene Transcripts in Tissues or Cells," *Anal. Biochem.*, 188:349–355 (1990).

Khrapko et al., "An Oligonucleotide hybridization approach to DNA sequencing," *FEBS Lett.*, 256(1,2):118–122 (1989).

Kievits et al., "Rapid subchromosomal localization of cosmids by nonradioactive in situ hybridization," *Cytogenetics Cell Genetics*, 53(2–3):134–136 (1990).

Kimura et al., "An Immobilized Enzyme Membrane Fabrication Method using an Ink Jet Nozzle," *Biosensors*, 4:41–52 (1988).

Kimura et al., "An Integrated SOF/FET Multi–Biosensor," *Sensors & Actuators*, 9:373–387 (1986).

Kitazawa et al., "In situ DNA–RNA hybridization using in vivo bromodeoxyuridine–labeled DNA probe," *Histochemistry*, 92:195–199 (1989).

Kleinfeld et al., "Controlled Outgrowth of Dissociated Neurons on Patterned Substrates," *J. Neurosci.*, 8(11):4098–4120 (1988).

Knight, P., "Materials and Methods/Microsequencers for Proteins and Oligosaccharides," *Bio/Tech.*, 7:1075–76 (1989).

Kohara et al., "The Physical Map of the Whole E. coli Chromosome: Application of a New Strategy for Rapid Analysis and Sorting of a Large Genomic Library," *Cell*, 50:495–508 (1987).

Krile et al., "Multiplex holography with chirp–modulated binary phase–coded reference–beam masks," *Applied Opt.*, 18(1):52–56 (1979).

Labat, I., "Subfragments as an informative characteristic of the DNA molecule—computer simulation," research report submitted to the University of Belgrade College of Nature Sciences and Mathematics, (1988).

Lainer et al., "Human Lymphocyte Subpopulations Identified by Using Three–Color Immunofluorescence and Flow Cytometry Analysis: Correlation of Leu–2, Leu–3, Leu–7, Leu–8, and Leu–11 Clee Surface Antigen Expression," *Journal of Immunology*, 132(1):151–156 (1984).

Lam et al., "A new type of synthetic peptide library for identifying ligand–binding activity," *Nature*, 354:82–84 (1991).

Laskey et al., "Messenger RNA prevalence in sea urchin embryos measured with cloned cDNAs," *PNAS*, 77(9):5317–5321 (1980).

Lee et al., "synthesis of a Polymer Surface Containing Covalently Attached Triethoxysilane Functionality: Adhesion to Glass," *Macromolecules*, 21:3353–3356 (1988).

Allister et al., "Labelling oligonucleotides to high specific activity (I)," *Nuc. Acids Res.*, 17(12):4605–4610 (89).

Frischauf et al., "Phage Vectors—EMBL Series," *Meth. Enzymology*, 153:103–115 (1987).

Levy, M.F., "Preparing Additive Printed Circuits," *IBM Tech. Discl. Bull.*, 9(11):1473 (1967).

Lichter et al., "High–Resolution Mapping of Human Chromosome 11 by in Situ hybridization with Cosmid Clones," *Science*, 247:64–69 (1990).

Lichter et al., "Fluorescence in situ hybridization with Alu and L1 polymerase chain reaction probes for rapid characterization of human chromosomes in hybrid cell lines," *PNAS*, 87:6634–6638 (1990).

Lichter et al., "Rapid detection of human chromosome 21 aberrations by in situ hybridization," *PNAS*, 85:9664–9668 (1988).

Lichter et al., "Is non–isotopic in situ hybridization finally coming of age," *Nature*, 345:93–94 (1990).

Lieberman et al., "A Light source Smaller Than the Optical Wavelength," *Science*, 247:59–61 (1990).

Lipshutz et al., "Using Oligonucleotide Probe Arrays To Access Genetic Diversity," *BioTech.*, 19(3):442–7 (1995).

Liu et al., "Sequential Injection Analysis in Capillary Format with an Electroosmotic Pump," *Talanta*, 41(11):1903–1910 (1994).

Lockhart et al., "Expression monitoring by hybridization to high–density oligonucleotide arrays," *Nat. Biotech.*, 14:1675–1680 (1996).

Logue et al., "General Approaches to Mask Design for Binary Optics," SPIE, 1052:19–24 (1989).

Loken et al., "three–color Immunofluorescence Analysis of Leu Antigens on Human Peripheral Blood Using Two Lasers on a Fluorescence–Activated Cell Sorter," *Cymoetry*, 5:151–158 (1984).

Love et al., "Screening of γ Library for Differentially Expressed Genes Using in Vitro Transcripts," *Anal. Biochem.*, 150:429–441 (1985).

Lowe, C.R., "Biosensors," *Trends in Biotech.*, 2:59–65 (1984).

Lowe, C.R., "An Introduction to the Concepts and Technology of Biosensors," *Biosensors*, 1:3–16 (1985).

Lowe, C.R., Biotechnology and Crop Improvement and Protection, BCPC Publications, pp. 131–138 (1986).

Lowe et al., "Solid–Phase Optoelectronic Biosensors," *Methods in Enzymology*, 137:338–347 (1988).

Lowe, C.R., "Biosensors," *Phil. Tran. R. Soc. Lond.*, 324:487–496 (1989).

Lu et al., "Differential screening of murine ascites cDNA libraries by means of in vitro transcripts of cell–cycle–phase–specific cDNA and digital image processing," *Gene*, 86:185–192 (1990).

Lysov et al., "A new method for determining the DNA nucleotide sequence by hybridization with oligonucleotides," *Doklady Biochem.*, 303(1–6):436–438 (1989).

Lysov et al., "DNA Sequencing by Oligonucleotide Hybridization," First International Conference on Electrophoresis, Supercomputing and the Human Genome, Apr. 10–13, 1990 p. 157.

MacDonald et al., "A Rapid ELISA for Measuring Insulin in a Large Number of Research Samples," *Metabolism*, 38(5):450–452 (1989).

Manz et al., "Miniaturized Total Chemical Analysis Systems: a Novel Concept for Chemical Sensing," *Sensors and Actuators*, B1:244–248 (1990).

Manz et al., "Micromachining of monocrystalline silicon and glass for chemical analysis systems, A look into next century's technology or just a fashionable craze?," *Trends in Analytical Chem.*, 10(5):144–149 (1991).

Manz et al., "Planar chips technology for minaturization and integration of separation techniques into monitoring systems, Capillary electrophoresis on a chip," *J. Chromatography*, 593:253–258 (1992).

Manz et al., "Planar Chips Technology for Miniaturization of Separation Systems: A Developing Perspective in Chemical Monitoring," chapter 1, 1–64 (1993).

Manz et al., "Electroosmotic pumping and electrophoretic separations for minaturized chemical analysis systems," *J. Micromech. Microeng.*, 4:257–265 (1994).

Masiakowski et al., "Cloning of cDNA sequences of hormone–regulated genes from the MCF–7 human breast cancer cell line," *Nuc. Acids Res.*, 10(24):7895–7903 (1982).

Matsumoto et al., "Preliminary Investigation of Micropumping Based on Electrical Control of Interfacial Tension," *IEEE*, pp. 105–110 (1990).

Matsuzawa et al., "Containment and growth of neuroblastoma cells on chemically patterned substrates," *J. Neurosci. Meth.*, 50:253–260 (1993).

McCray et al., "Properties and Uses of Photoreactive Caged Compounds," *Ann. Rev. Biophys. Biophys. Chem.*, 18:239–270 (1989).

McGall et al., "The Efficiency of Light–Directed Synthesis of DNA Arrays on Glass Substrates," *J. American Chem. Soc.*, 119(22):5081–5090 (1997).

McGillis, VLSI Technology, Sze, eds., Chapter 7, "Lithography," pp. 267–301 (1983).

McMurray, J.S., "Solid Phase Synthesis of a Cyclic Peptide Using Fmoc Chemistry," *Tetrahedron Letters*, 32(52):7679–7682 (1991).

Meinkoth et al., "Review: Hybridization of Nucleic Acids Immobilized on solid Supports," *Analytical Biochem.*, 138:267–284 (1984).

Melcher et al., "Traveling–Wave Bulk Electroconvection Induced across a Temperature Gradient," *Physics of Fluids*, 10(6):1178–1185 (1967).

Merrifield, R.B., "Solid Phase peptide Synthesis. I. The Synthesis of a Tetrapeptide," *J.Am.Chem.Soc.*, 85:2149–2154 (1963).

Michiels et al., "Molecular approaches to genome analysis: a strategy for the construction of ordered overlapping clone libraries," *CABIOS*, 3(3):203–10 (1987).

Mirzabekov, A.D., "DNA sequencing by hybridization—a megasequencing method and a diagnostic tool?," *TIBTECH*, 12:27–32 (1994).

Monaco et al., "Human Genome Linking with Cosmids and Yeast Artificial Chromosomes", abstract from CSHS, p. 50, (1989).

Morrison et al., "Solution–Phase Detection of Polynucleotides Using Interacting Fluorescent Labels and Competitive Hybridization," *Anal. Biochem.*, 183:231–244 (1989).

Mutter et al., "Impact of Conformation on the Synthetic Strategies for Peptide Sequences," pp. 217–228 from Chemistry of Peptides and Proteins, vol. 1, Proceedings of the Third USSR–FRG Symp., in USSR (1982).

Nakamori et al., "A Simple and Useful Method for Simultaneous Screening of Elevated Levels of Expression of a Variety of Oncogenes in Malignant Cells," *Jpn. J. Cancer Res.*, 79:1311–1317 (1988).

Nederlof et al., "Multiple Fluorescence In Situ Hybridization," *Cytometry*, 11:126–131 (1990).

Nyborg, W., "Acoustic Streaming," chapter 11 pp. 265–329 from Physical Acoustics, Principles and Methods, Mason, eds., vol. II, part B, Academic Press, New York and London (1965).

Ocvirk et al., "High Performance Liquid Chromatography Partially Integrated onto a Silicon Chip," *Analyt. Meth. Instrumentation*, 2(2):74–82 (1995).

Ohtsuka et al., "Studies on transfer ribonucleic acids and related compounds. IX Ribonucleic oligonucleotide synthesis using a photosensitive 0–nitrobenzyl protection at the 2'–hydroxyl group," *NucAcids.Res.*, 1(10):1351–1357 (1974).

Olefirowicz et al., "Capillary Electrophoresis for Sampling Single Nerve Cells," *Chimia*, 45(4):106–108 (1991).

Patchornik et al., "Photosensitive Protecting Groups," *J.Am-.Chem.Soc.*, 92(21):6333–6335 (1970).

Patent Abstracts of Japan from EPO, Abst. 13:557, JP 1–233 447 (1989).

Pease et al., "Light–generated oligonucleotide arrays for rapid DNA sequence analysis," *PNAS*, 91:5022–26 (1994).

Pevzner, P.A., "1–Tuple DNA Sequencing: Computer Analysis," *J. Biomol. Struct. Dynam.*, 7(1):63–69 (1989).

Pfahler et al., "Liquid Transport in Micron and Submicron Channels," *Sensors and Actuators*, A21–A23:431–4 (90).

Pidgeon et al., "Immobilized Artificial Membrane Chromatography: Supports Composed of Membrane Lipids," *Anal. Biochem.*,176:36–47 (89).

Pillai, V.N., "Photoremovable Protecting Groups in Organic Synthesis," *Synthesis*, pp. 1–26 (1980).

Pillai et al., "3–Nitro–4–Aminomethylbenzoylderivative von Polyethylenglykolen: Eine neue Klasse von Photosensitiven loslichen Polymeren Tragern zur Synthese von C–terminalen Peptidamiden," *Tetrah. ltr.*, # 36 pp. 3409–3412 (1979).

Pillai et al., "Synthetic Hydrophilic Polymers, Biomedical and Chemical Applications," *Naturwissenschaften*, 68:558–566 (1981).

Pirrung et al., "Proofing of Photolithographic DNA Synthesis with 3',5'–Dimethoxybenzoinyloxycarbonyl–Protected Deoxynucleoside Phosphoramidites," *J. Org. Chem.*, 63(2):241–246 (1998).

Pirrung et al., "Comparison of Methods for Photochemical Phosphoramidite–Based DNA Synthesis," *J. Org. Chem.*, 60:6270–6276 (1995).

Ploax et al., "Cyclization of peptides on a solid support," *Int. J. Peptide Protein Research*, 29:162–169 (1987).

Polsky–Cynkin et al., "Use of DNA Immobilized on Plastic and Agarose Supports to Detect DNA by Sandwich Hybridization," *Clin. Chem.*, 31(9):1428–1443 (1985).

Poustka et al., "Molecular Approaches to Mammalian Genetics," Cold Spring Harbor Symposia on Quantitative Biology, 51:131–139 (1986).

Purushothaman et al., "Synthesis of 4,5–diarylimidazoline–2–thiones and their photoconversion to bis(4,5–diarylimidzaol–2–yl) sulphides," *Ind. J. Chem.*, 29B:18–21 (1990).

Quesada et al., "High–Sensitivity DNA Detection with a Laser–Exited Confocal Fluorescence Gel Scanner," *Biotechniques*, 10:616 (1991).

Reichmanis et al., *J. Polymer Sci. Polymer Chem. Edition*, 23:1–8 (1985).

Richter et al., "An Electrohydrodynamic Micropump," *IEEE*, pp. 99–104 (1990).

Richter et al., "Electrohydrodynamic Pumping and Flow Measurement," *IEEE*, pp. 271–276 (1991).

Richter et al., "A Micromachined electrohydrodynamic (EHD) pump," *Sensors and Actuators*, A29:159–168 (91).

Robertson et al., "A General and Efficient Route for Chemical Aminoacylation of Transfer RNAs," *J. Am. Chem. Soc.*, 113:2722–2729 (1991).

Rodda et al., "The Antibody Response to Myoglobin–I. Systematic Synthesis of Myglobin Peptides Reveals Location and Substructure of Species–Dependent Continuous Antigenic Determinants," *Mol. Immunol.*, 23(6):603–610 (1986).

Rodgers, R.P., "Data Processing of Immunoassay Results," Manual of Clin. Lab. Immunol., 3rd ed., ch. 15, pp. 82–87 (1986).

Rose, D.J., "Free–solution reactor for post–column fluorescence detection in capillary zone electrophoresis," *J. Chromatography*, 540:343–353 (1991).

Rovero et al., "Synthesis of Cylic Peptides on solid Support," *Tetrahed. Letters*, 32(23):2639–2642 (1991).

Saiki et al., "Genetic analysis of amplified DNA with immobilized sequence–specific oligonucleotide probes," *PNAS*, 86:6230–6234 (1989).

Saiki et al., "Analysis of enzymatically amplified β–globin and HLA–DQα DNA with Allele–specific oligonucleotide probes," *Nature*, 324:163–166 (1986).

Scharf et al., "HLA class II allelic variation and susceptibility to pemphigus vulgaris," *PNAS*, 85(10):3504–3508 (1988).

Schuup et al., "Mechanistic Studies of the Photorearrangement of o–Nitrobenzyl Esters," *J. Photochem.*, 36:85–97 (1987).

Seiler et al., "Planar Glass Chips for Capillary Electrophoresis: Repetitive Sample Injection, Quantitation, and Separation Efficency," *Anal. Chem.*, 65:1481–1488 (1993).

Seller et al., "Electroosmotic Pumping and Valveless Control of Fluid Flow within a Manifold of Capillaries on a Glass Chip," Anal. Chem., 66:3485–3491 (1994).

Sheldon et al., "Matrix DNA Hybridization," *Clinical Chemistry*, 39(4):718–719 (1993).

Shin et al., "Dehydrooligonpeptides. XI. Facile Synthesis of Various Kinds of Dehydrodi– and tripeptides, and Dehydroenkephalins Containing Tyr Residue by Using N–Carboxydehydrotyrosine Anhydride," *Bull. Chem. Soc. Jpn.*, 62:1127–1135 (1989).

Sim et al., "Use of a cDNA Library for Studies on Evolution and Development Expression of the Chorion Multigene Families," *Cell*, 18:1303–1316 (1979).

Smith et al., "A Novel Method for Delineating Antigenic Determinants: Peptide Synthesis and Radioimmunoassay Using the Same Solid Support," *Immunochemistry*, 14:565–568 (1977).

Southern et al., "Report on the Sequencing by Hybridization Workshop," *Genomics*, 13:1378–1383 (1992).

Southern et al., "Oligonucleotide hybridisations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesized in situ," *Nuc. Acids Res.*, 20(7):1679–1684 (1992).

Southern et al., "Analyzing and Comparing Nucleic Acid Sequences by Hybridization to Arrays of Oligonucleotides: Evaluation Using Experimental Models," Genomics, 13:1008–10017 (1992).

Stemme et al., "A valveless diffuser/nozzle–based fluid pump," *Sensors and Actuators*, A39:159–167 (1993).

Stryer, L., "DNA Probes and Genes Can be Synthesized by Automated Solid–Phase Methods," from *Biochemistry*, Third Edition, published by W.H. Freeman & Co., (1988).

Stuber et al., "Synthesis and photolytic cleavage of bovine insulin B22–30 on a nitrobenzoylglycyl–poly (ethylene glycol) support," *Int. J. Peptide Protein Res.*, 22(3):277–283 (1984).

Sundberg et al., "Spatially–Addressable Immobilization of Macromolecules on Solid Supports," *J. Am. Chem. Soc.*, 117(49):12050–12057 (1995).

Swedberg, S.A., "Use of non–ionic and zwitterionic surfactants to enhance selectivity in high–performance capillary electrophoresis, An apparant micellar electrokinetic capillary chromatography mechanism," *J. Chromatography*, 503:449–452 (1990).

Titus et al., "Texas Red, a Hydrophilic, red–emitting fluorophore for use with fluorescein in dual parameter plow microfluorometric and fluorescence microscopic studies," *J. Immunol. Meth.*, 50:193–204 (1982).

Tkachuk et al., "Detection of bcr–abl Fusion in chronic Myelogeneous Leukemia by in situ Hybridization," *Science*, 250:559–562 (90).

Trzeciak et al., "Synthesis of 'Head–to–Tail' Cyclized Peptides on Solid Support by FMOC Chemistry," *Tetrahed. Letters*, 33(32):4557–4560 (1992).

Tsien et al., "Control of Cytoplasmic Calcium with Photolabile Tetracarboxylate 2–Nitrobenzhydrol Chelators," *Biophys. J.*, 50:843–853 (1986).

Tsutsumi et al., "Expression of L– and M–Type Pyruvate Kinase in Human Tissues," *Genomics*, 2:86–89 (1988).

Turchinskii et al., "Multiple Hybridization in Genome Analysis, Reaction of Diamines and Bisulfate with Cytosine for Introduction of Nonradioactive labels Into DNA," *Molecular Biology*, 22:1229–1235 (1988).

Turner et al., "Photochemical Activation of Acylated α–Thrombin," *J. Am. Chem. Soc.*, 109:1274–1275 (1987).

Urdea et al., "A novel method for the rapid detection of specific nucleotide sequences in crude biological sample without blotting or radioactivity; application to the analysis of hepatitis B virus in human serum," *Gene*, 61:253–264 (1987).

Urdea et al., "A comparison of non–radioisotope hybridization assay methods using fluorescent, chemiluminescent and enzyme labeled synthetic oligodeoxyribonucleotide probes," *Nuc. Acids Res.*, 16(11):4937–4956 (1988).

Van der Voort et al., "Design and Use of a Computer Controlled Confocal Microscopic for Biological Applications," *Scanning*, 7(2):66–78 (1985).

Van Hijfte et al., "Intramolecular 1,3–Diyl Trapping Reactions. A Formal Total Synthesis of –Coriolin," J. Organic Chemistry, 50:3942–3944 (1985).

Veldkamp, W.B., "Binary optics: the optics technology of the 1990s," CLEO 90, vol. 7, paper # CMG6 (1990).

Verlaan–de Vries et al., "A dot–blot screening procedure for mutated ras oncogenes using synthetic oligodeoxynucleotides," *Gene*, 50:313–320 (1986).

Verpoorte et al., "Three–dimensional micro flow manifolds for miniaturized chemical analysis systems," *J. Micromech. Microeng.*, 4:246–256 (1994).

Volkmuth et al., "DNA electrophoresis in microlithographic arrays," *Nature*, 358:600–602 (1992).

Voss et al., "The immobilization of oligonucleotides and their hybridization properties," *Biochem. Soc. Transact.*, 16:216–217 (1988).

Walker et al., "Photolabile Protecting Groups for an Acetylcholine Receptor Ligand. Synthesis and Photochemistry of a New Class of o–Nitrobenzyl Derivatives and their Effects on Receptor Function," *Biochemistry*, 25:1799–1805 (1986).

Wallace et al., "Hybridization of synthetic oligodeoxyribonucleotides to Φχ 174 DNA: the effect of single base pair mismatch," *Nuc. Acids Res.*, 11(6):3543–3557 (1979).

Washizu et al., "Handling Biological Cells Using a Fluid Integrated Circuit," *IEEE Transactions Industry Applications*, 26(2):352–358 (1990).

Werner et al., "Size–Dependent Separation of Proteins Denatured in SDS by Capillary Electrophoresis Using a Replaceable Sieving Matrix," *Anal. Biochem.*, 212:253–258 (1993).

White et al., "An Evaluation of Confocal Versus Conventional Imaging of Biological Structures by Fluorescence Light Microscopy," *J. Cell Biol.*, 105(1):41–48 (1987).

Widacki et al., "Biochemical Differences in Qa–2 Antigens Expressed by Qa–2+,6+ and Qa–2a+,6– Strains. Evidence for Differential Expression of the Q7 and Q9 Genes," *Mol. Immunology*, 27(6):559–570 (1990).

Wilcox et al., "Synthesis of Photolabile 'Precursors' of Amino Acid Neurotransmitters," *J. Org. Chem.*, 55:1585–1589 (1990).

Wilding et al., "PCR in a Silicon Microstructure," *Clin. Chem.*, 40(9):1815–1818 (1994).

Wilding et al., "Manipulation and Flow of Biological Fluids in Straight Channels Micromachined in Silicon," *Clin. Chem.*, 40(1):43–47 (1994).

Wittman–Liebold, eds., Methods in Protein Sequence Analysis, from Proceedings of 7th Int'l Conf., Berlin, Germany, Jul. 3–8, 1988, table of contents, pp. xi–xx* (1989).

Woolley et al., "Ultra–high–speed DNA fragment separations using microfabricated capillary array electrophoresis chips," *PNAS*, 91:11348–11352 (1994).

Wu et al., "Synthesis and Properties of Adenosine–5'–triphosphoro–γ–5–(5–sulfonic acid)naphthyl Ethylamidate: A Fluorescent Nucleotide Substrate for DNA–Dependent RNA Polymerase from *Escherichia coli,*" *Arch. Biochem. Biophys.*, 246(2):564–571 (1986).

Wu et al., "Laboratory Methods, Direct Analysis of Single Nucleotide Variation in Human DNA and RNA Using In Situ Dot Hybridization," *DNA*, 8(2):135–142 (1989).

Yamamoto et al., "Features and applications of the laser scanning microscope," *J. Mod. Optics*, 37(11):1691–1701 (1990).

Yarbrough et al., "Synthesis and Properties of Fluorescent Nucleotide Substrates for DNA–dependent RNA Polymerases," *J. Biol. Chem.*, 254(23):12069–12073 (1979).

Yosomiya et al., "Performance, Glass fiber Having Isocyanate Group on the Surface. Preparation and Reaction with Amino Acid," *Polymer Bulletin*, 12:41–48 (1984).

Young, W.S., "Simultaneous Use of Digoxigenin– and Radiolabeled Oligodeoxyribonucleotide Probes for Hybridization Histochemistry," *Neuropeptides*, 13:271–275 (1989).

Yue et al., "Minaiture Field–Flow Fractionation System for Analysis of Blood Cells," *Clin. Chem.*, 40(9):1810–1814 (1994).

Zehavi et al., "Light–Sensitive Glycosides. I. 6–Nitroveratryl β–D–Glucopyranoside and 2–Nitrobenzyl β–D–Glucopyranoside," *J. Org. Chem.*, 37(14):2281–2285 (1972).

Zengerle et al., "Transient measurements on miniaturized diaphragm pumps in microfluid systems," *Sensors and Actuators*, A46–47:557–561 (1995).

Ekins et al., "Fluorescence Spectroscopy and its Application to a New Generation of High Sensitivity, Multi–Microspot, Multianalyte, Immunoassay," *Clin. Chim. Acta*, 194:91–114 (1990).

Hames et al., *Nuclear acid hybridization, a practical approach*, cover page and table of contents (1985).

Mairanovsky, V. G., "Electro–Deprotection–Electrochemical Removal of Protecting Groups**," *Agnew. Chem. Int. Ed. Engl.*, 15(5):281–292 (1976).

Morita et al., "Direct pattern fabrication on silicone resin by vapor phase electron beam polymerization," *J. Vac. Sci. Technol.*, B1(4):1171–1173 (1983).

Munegumi et al., "thermal Synthesis of Polypeptides from N–Boc–Amino Acid (Aspartic Acid, β–Aminoglutaric Acid) Anhydrides, " *Chem. Letters*, pp. 1643–1646 (1988).

Sambrook, Molecular Cloning –A Laboratory Manual, publ. in 1989 (not included).

Semmelhack et al., "Selective Removal of Protecting Groups Using Controlled Potential Electrolysis," *J. Am. Chem. Society*, 94(14):5139–5140 (1972).

* cited by examiner

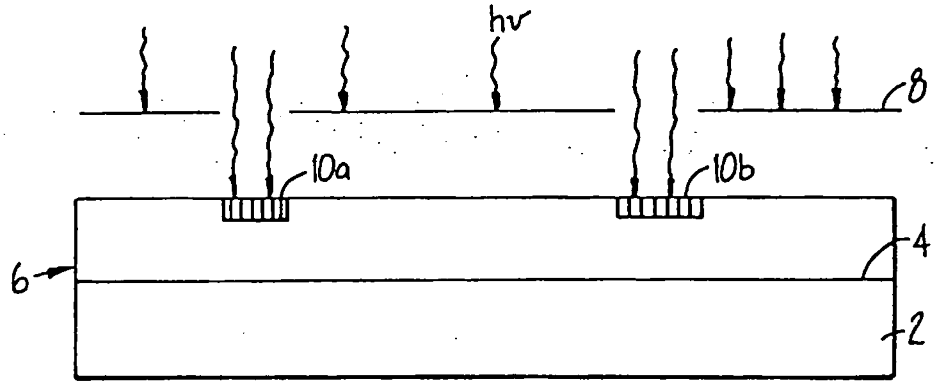


FIG._1.

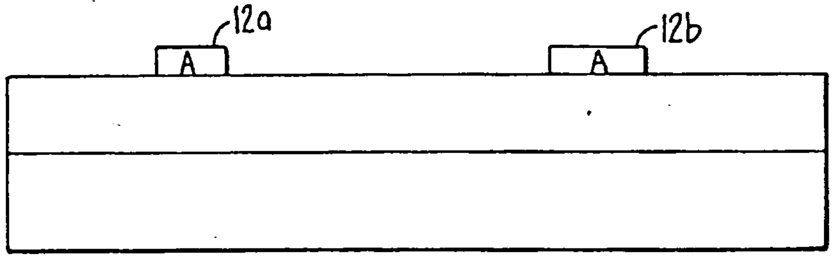


FIG._2.

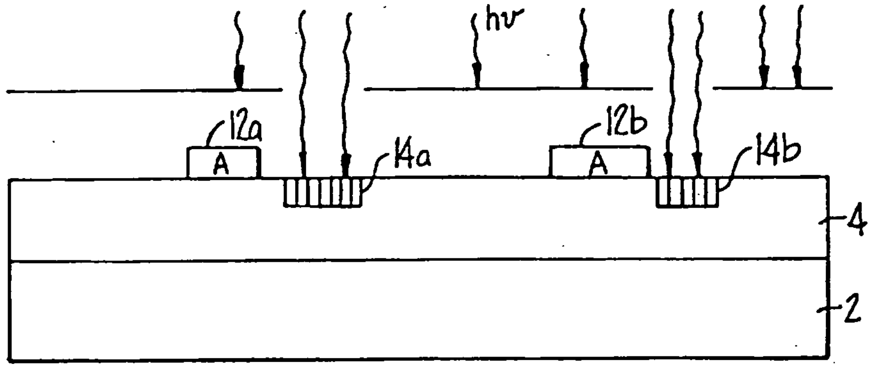


FIG._3.

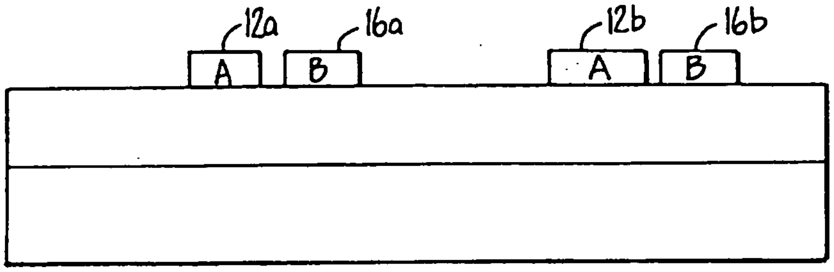


FIG._4.

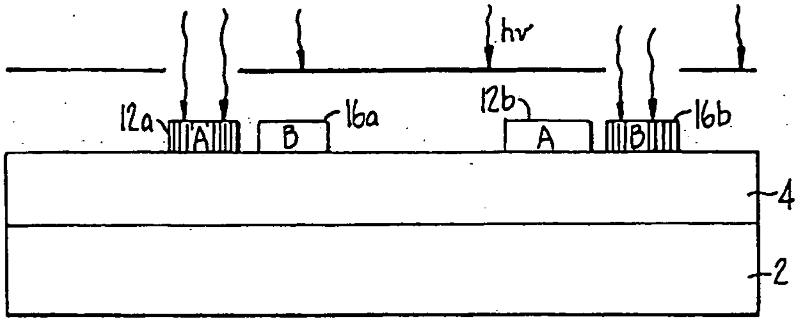


FIG._5.

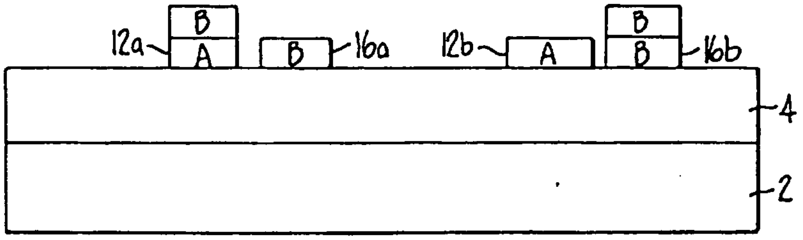


FIG._6.

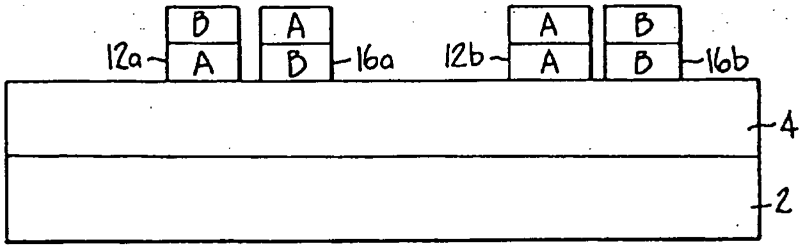


FIG._7.

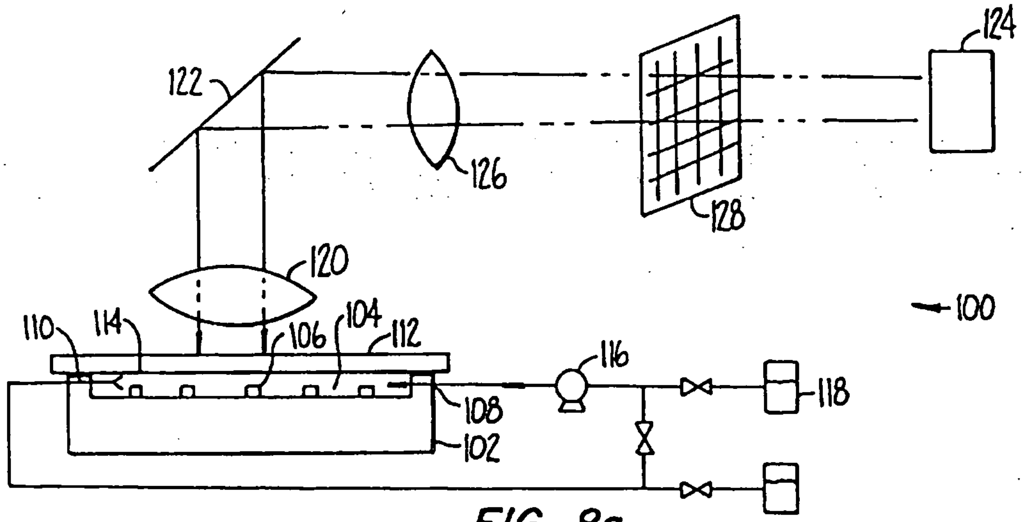


FIG._8a.

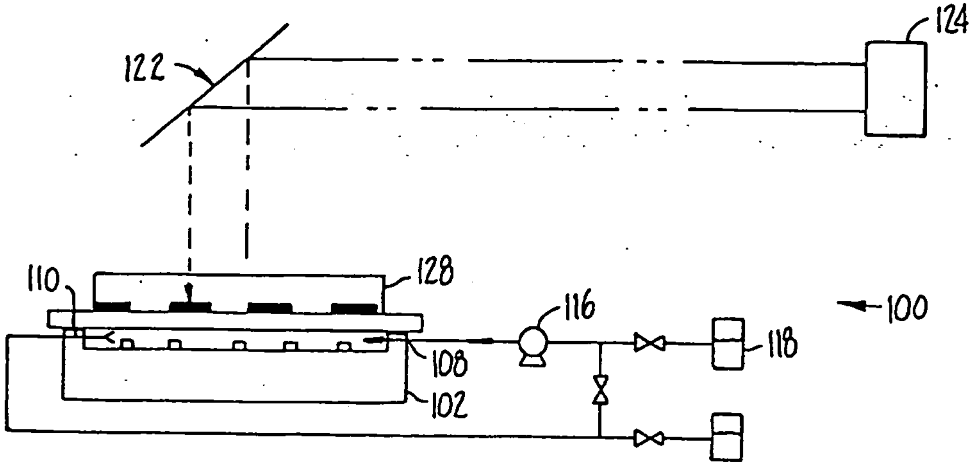


FIG._8b.

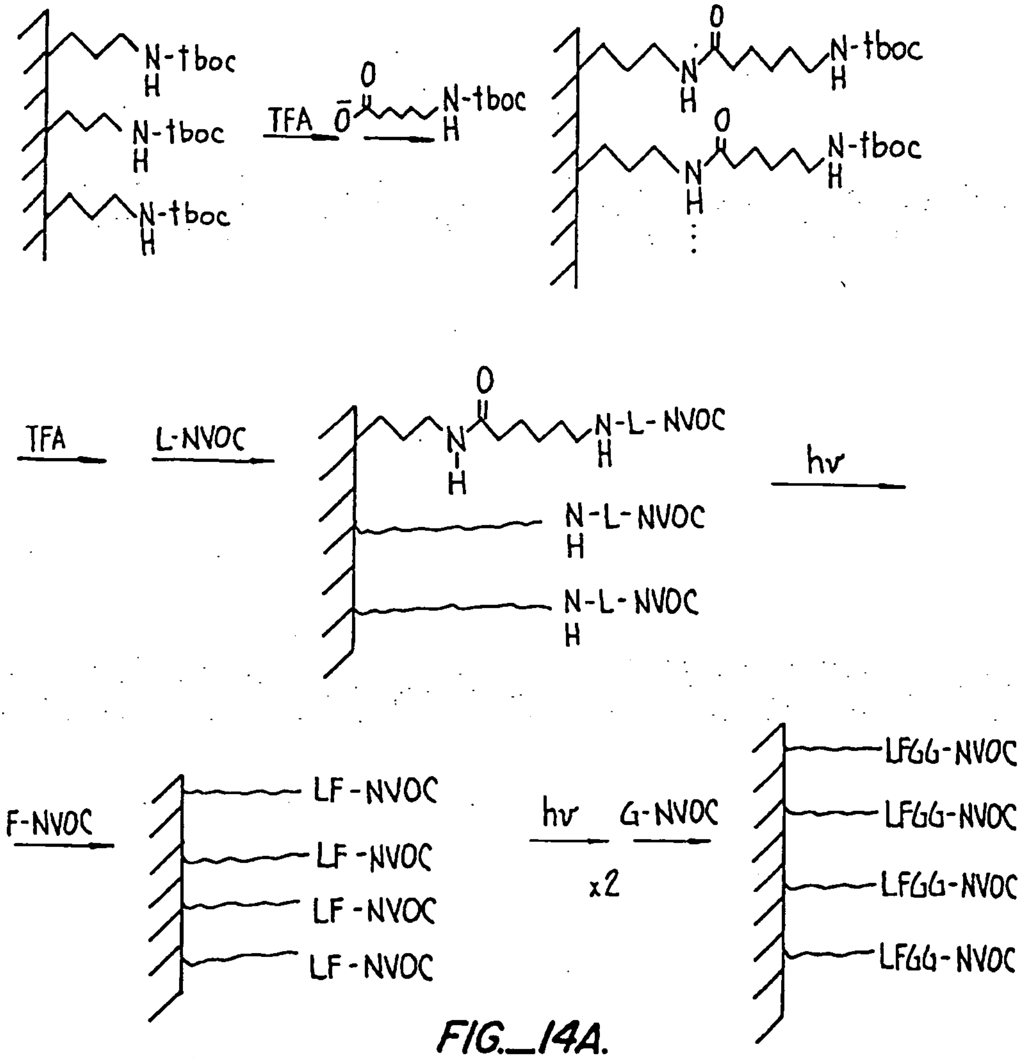


FIG._14A.

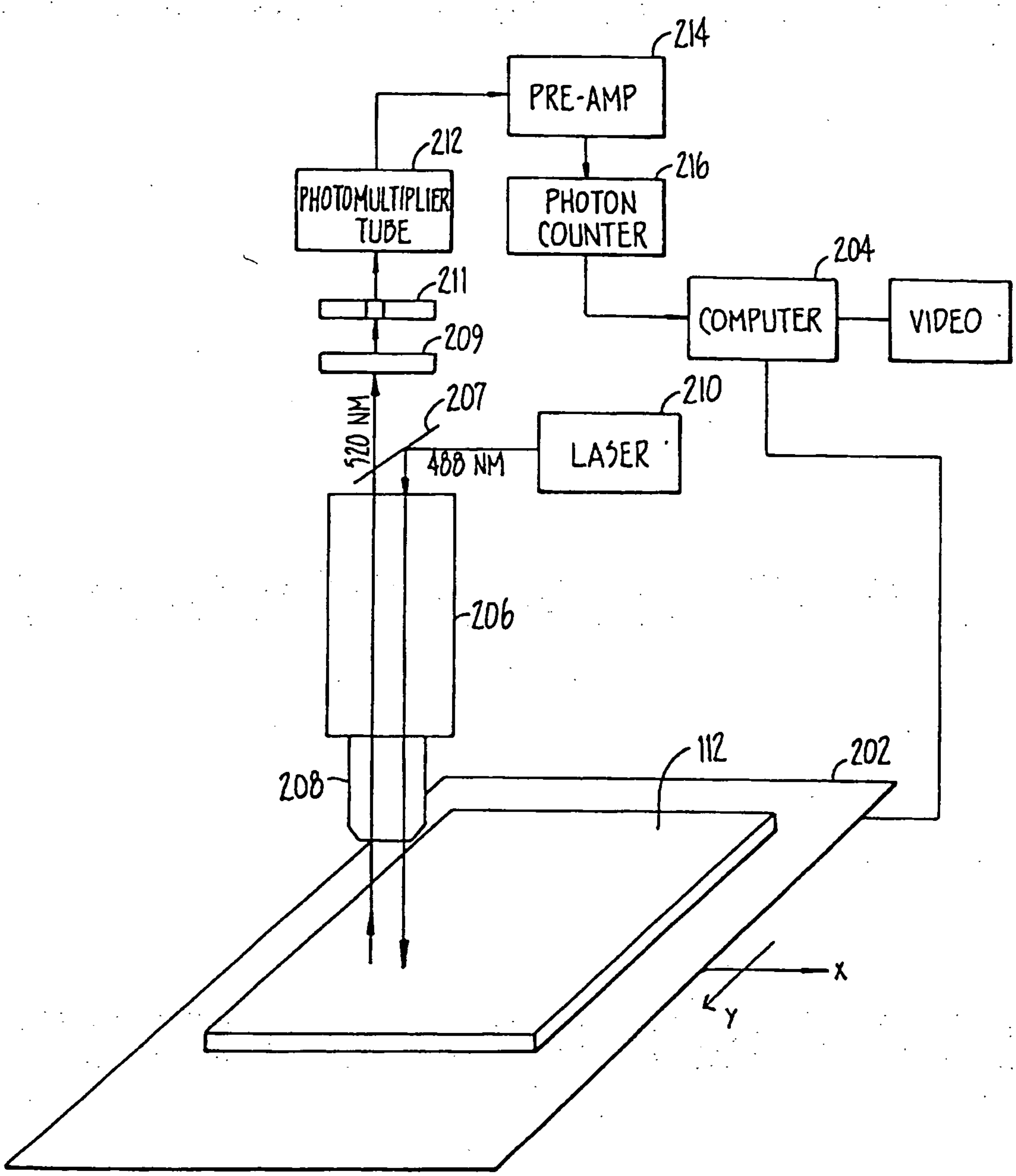

214
PRE-AMP
212
PHOTOMULTIPLIER TUBE
216
PHOTON COUNTER
204
COMPUTER
VIDEO
211
209
207
210
520 NM
488 NM
LASER
206
112
202
208
X
Y

FIG._9.

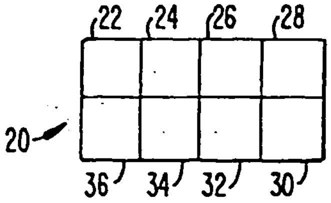


FIG._10A.

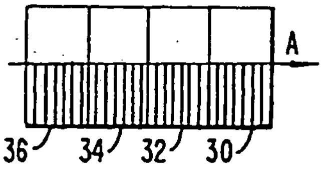


FIG._10B.

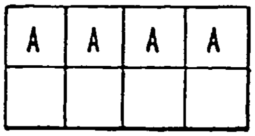


FIG._10C.

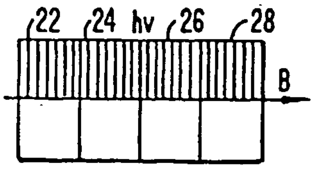


FIG._10D.

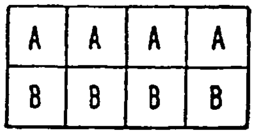


FIG._10E.

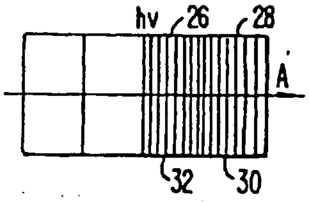


FIG._10F.

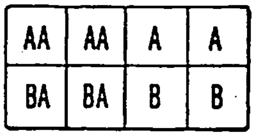


FIG._10G.

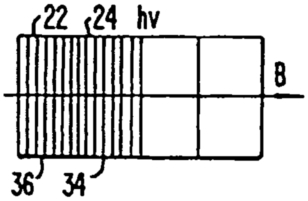


FIG._10H.

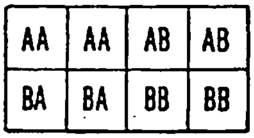


FIG._10I.

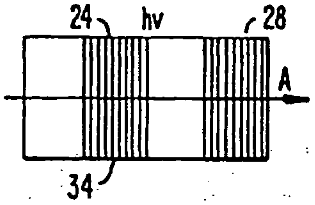


FIG._10J.

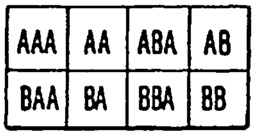


FIG._10K.

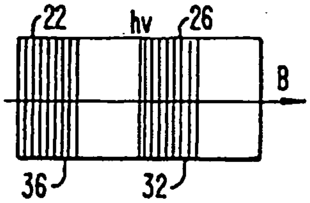


FIG._10L.

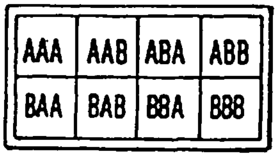


FIG._10M.

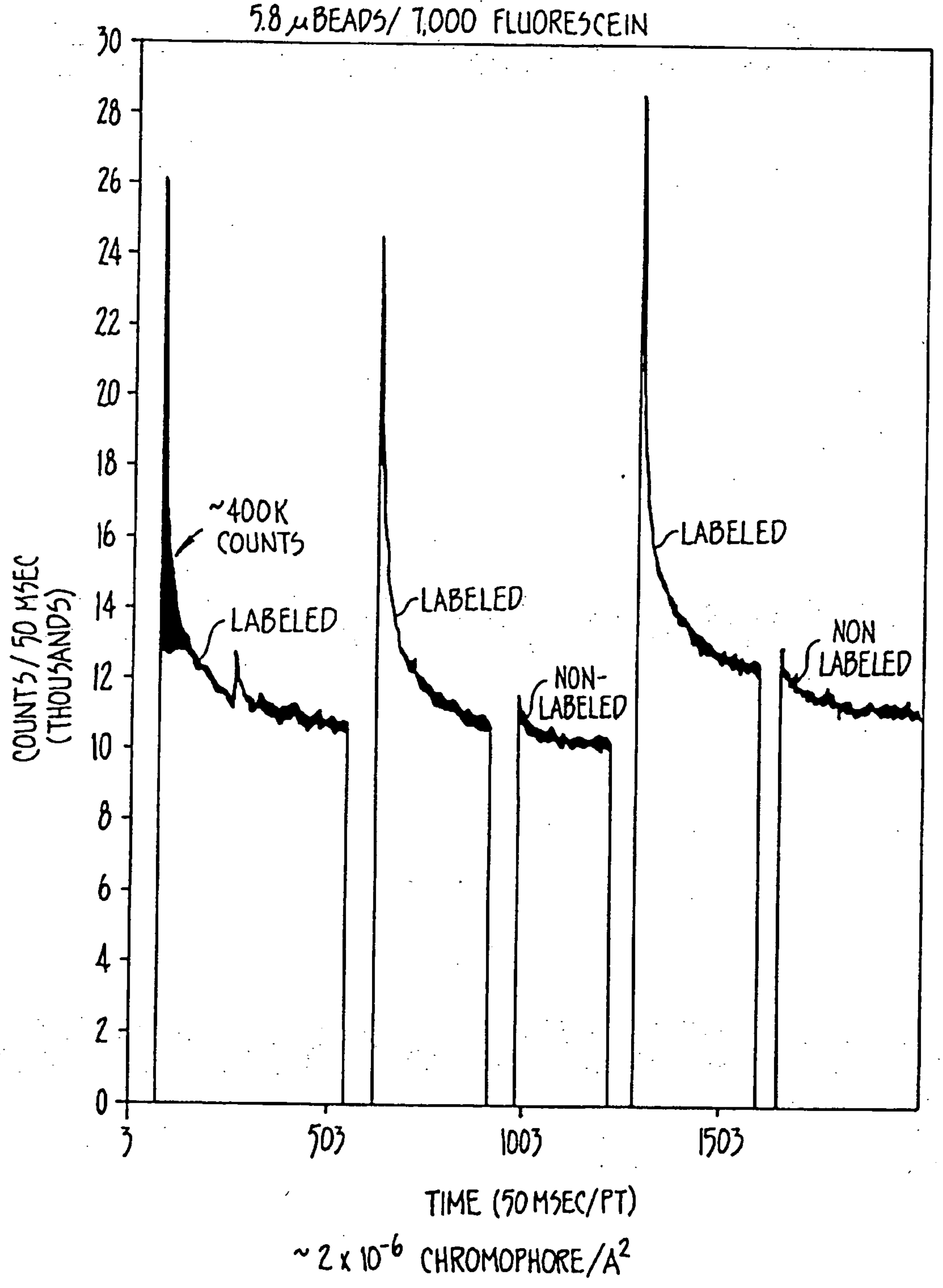

5.8 μBEADS/ 7,000 FLUORESCEIN
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
COUNTS / 50 MSEC
(THOUSANDS)
~400K
COUNTS
LABELED
LABELED
NON-
LABELED
LABELED
NON
LABELED
3
503
1003
1503
TIME (50 MSEC/PT)
~2 x 10^{-6} CHROMOPHORE/A^2

FIG._11A.

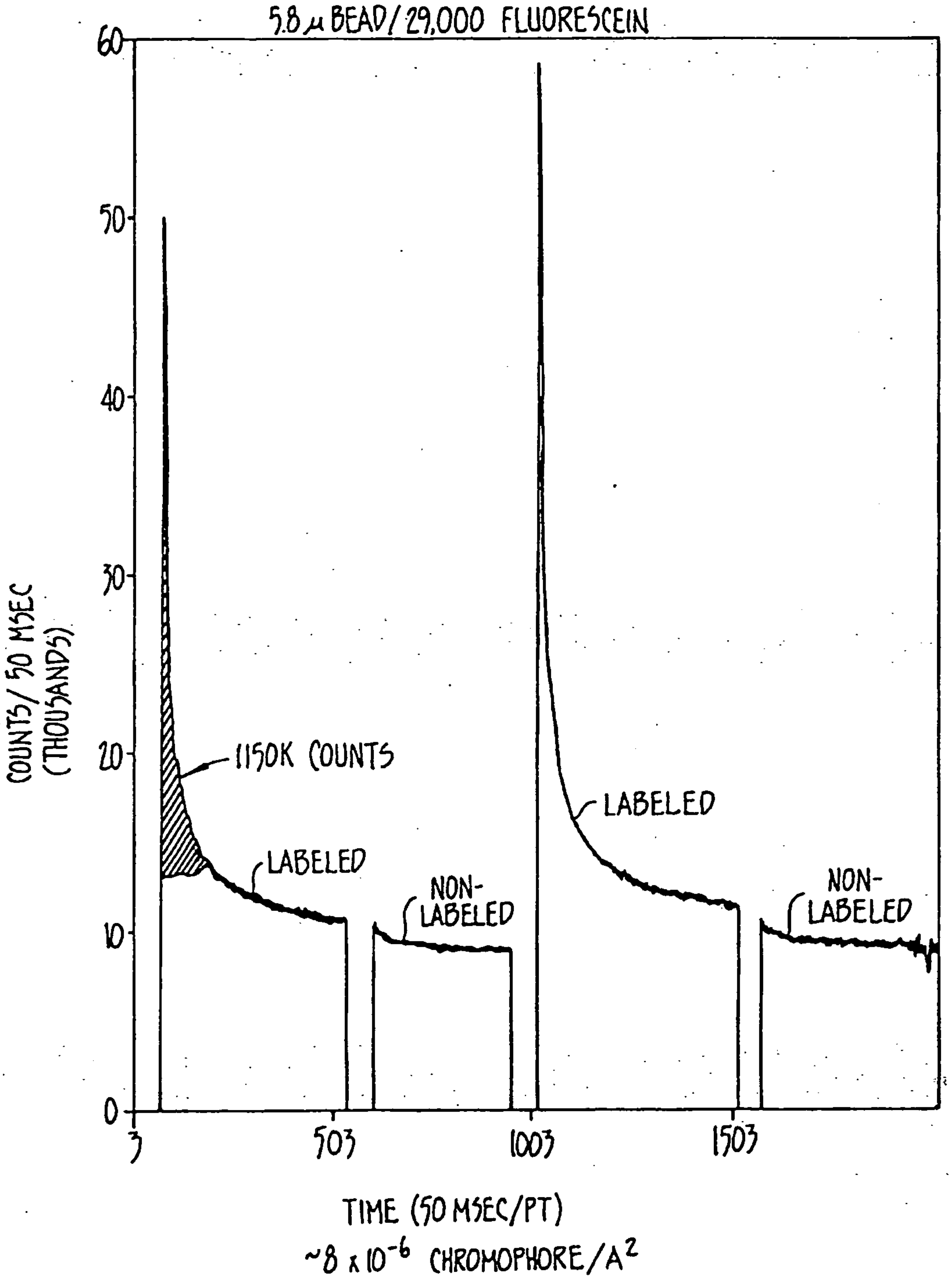

5.8μ BEAD/29,000 FLUORESCEIN
COUNTS/50 MSEC (THOUSANDS)
60
50
40
30
20
10
0
1150K COUNTS
LABELED
NON-LABELED
LABELED
NON-LABELED
3
503
1003
1503
TIME (50 MSEC/PT)
~8 x 10^-6 CHROMOPHORE/A^2

FIG._11B.

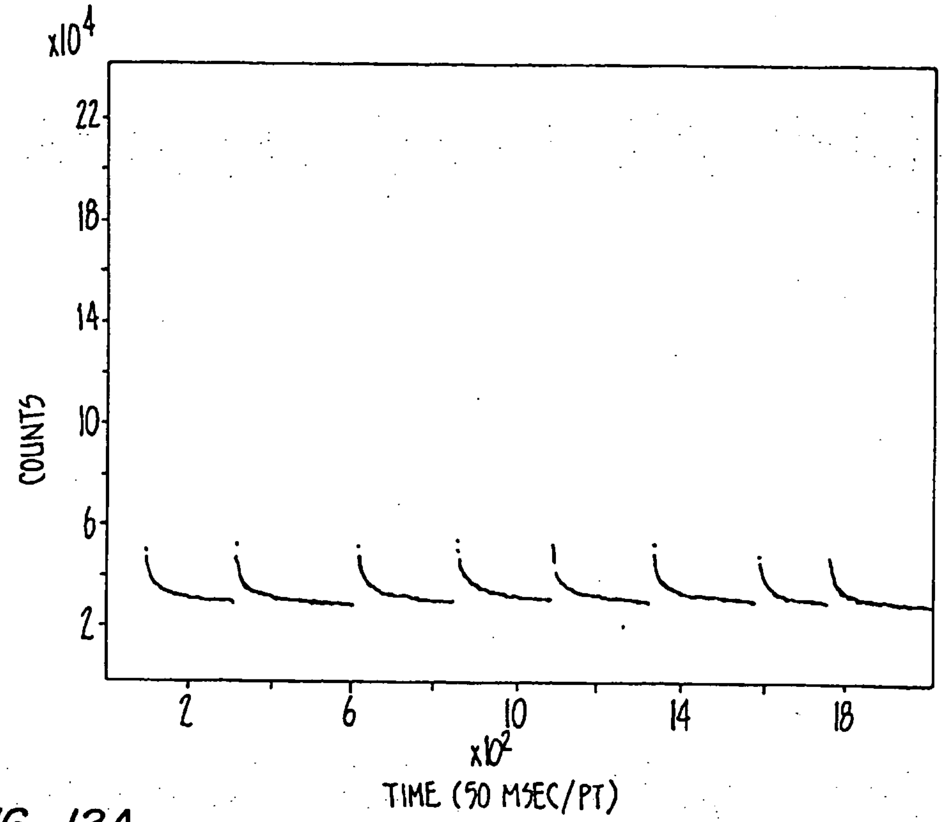


FIG._12A.

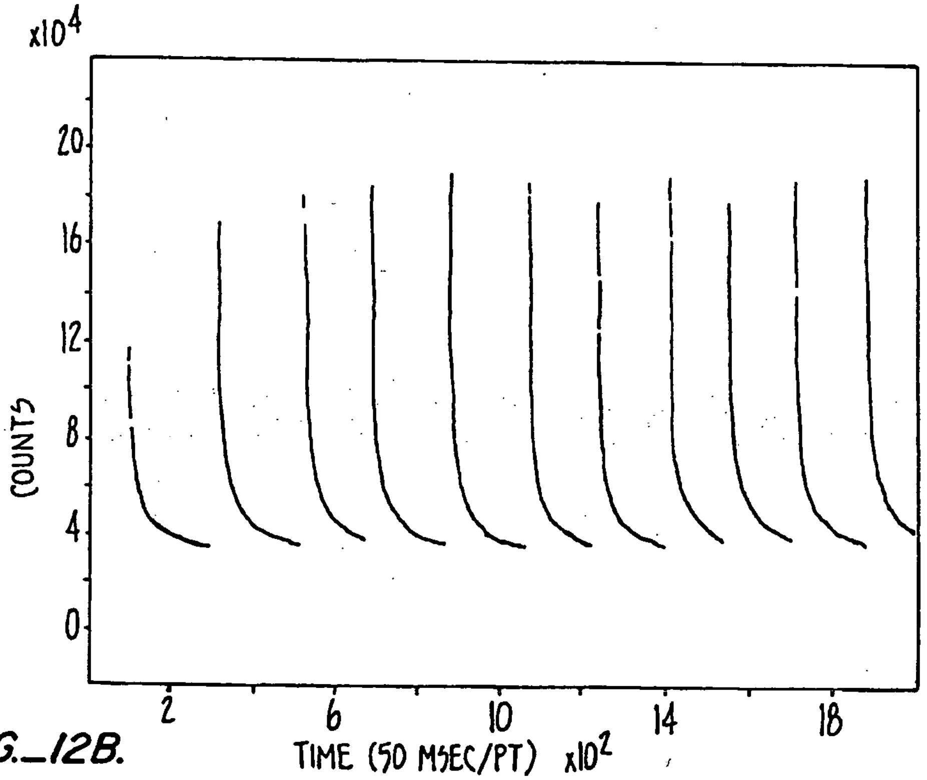
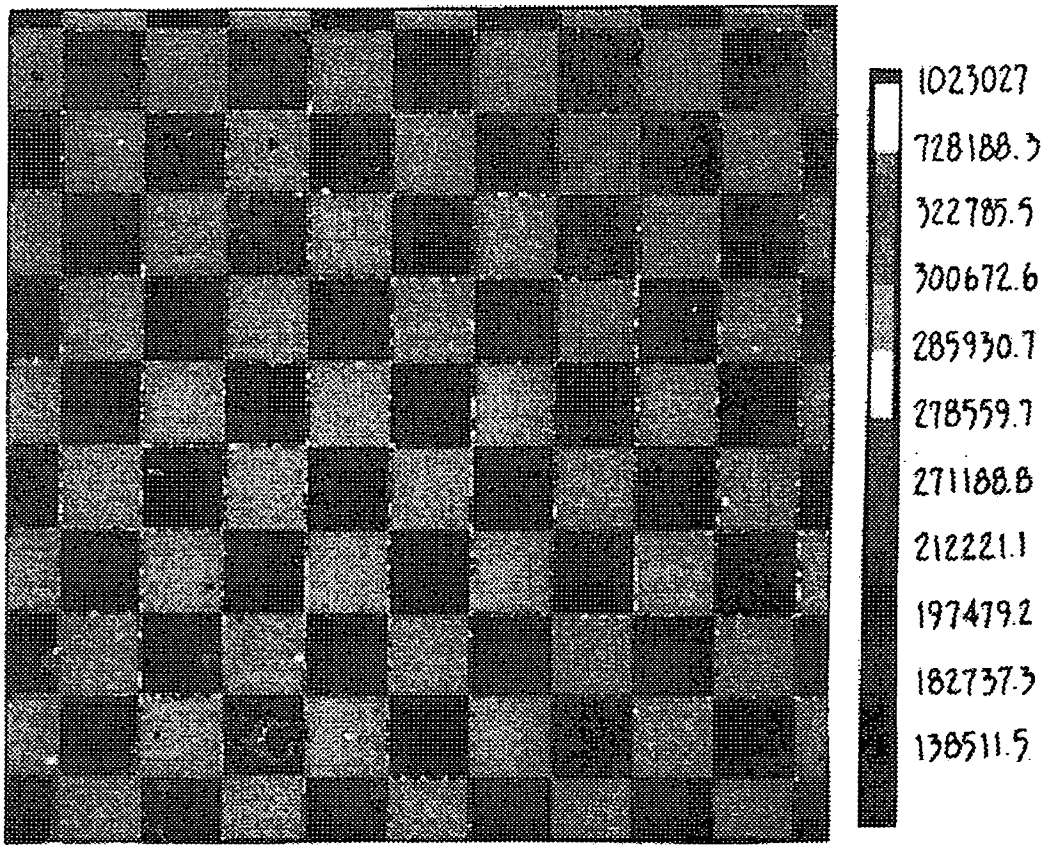
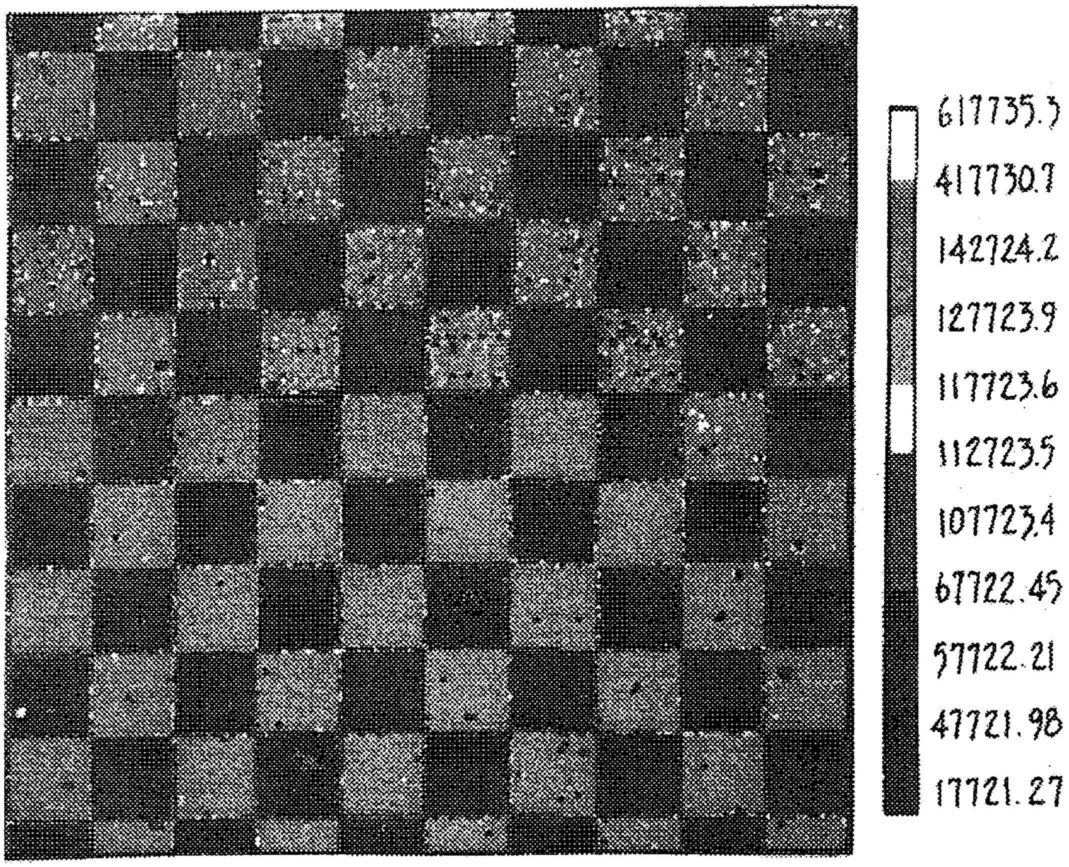


FIG._12B.



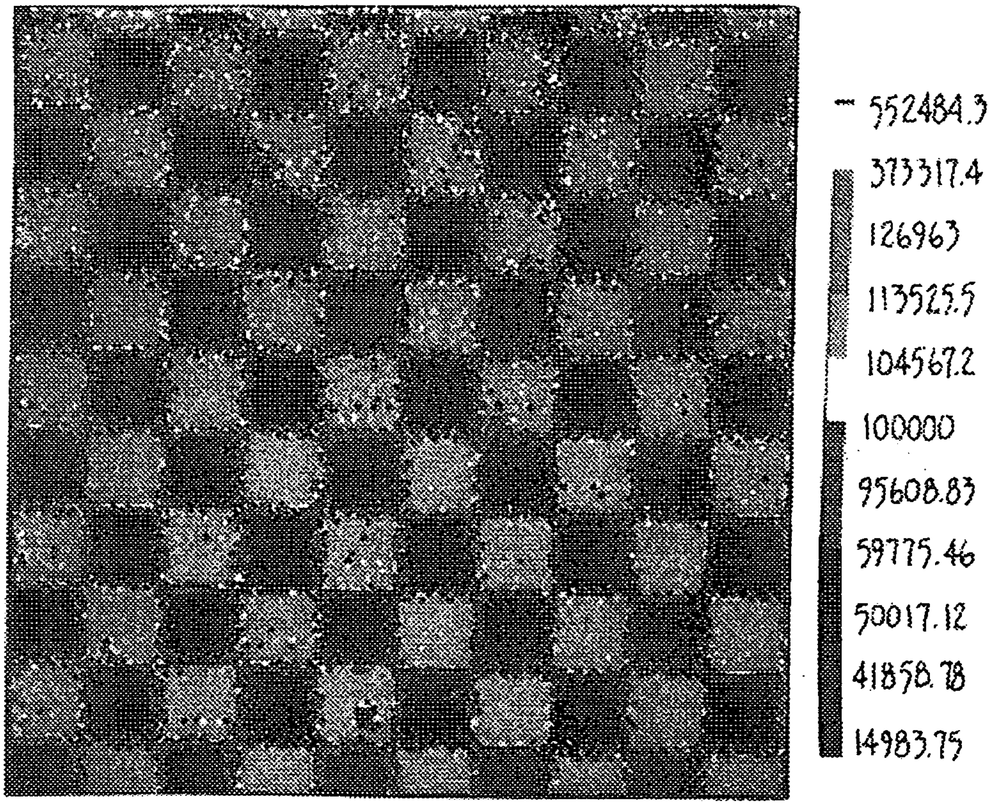
MEAN: 285930.7
VAR: 2.173242E+10
σ: 147419.2

FIG._13A.



MEAN: 117723.6
VAR: 1.000047E+10
σ: 100002.3

FIG._13B.



MEAN: 104567.2
VAR: 8.025189E+09
σ: 89583.42

FIG._13C.

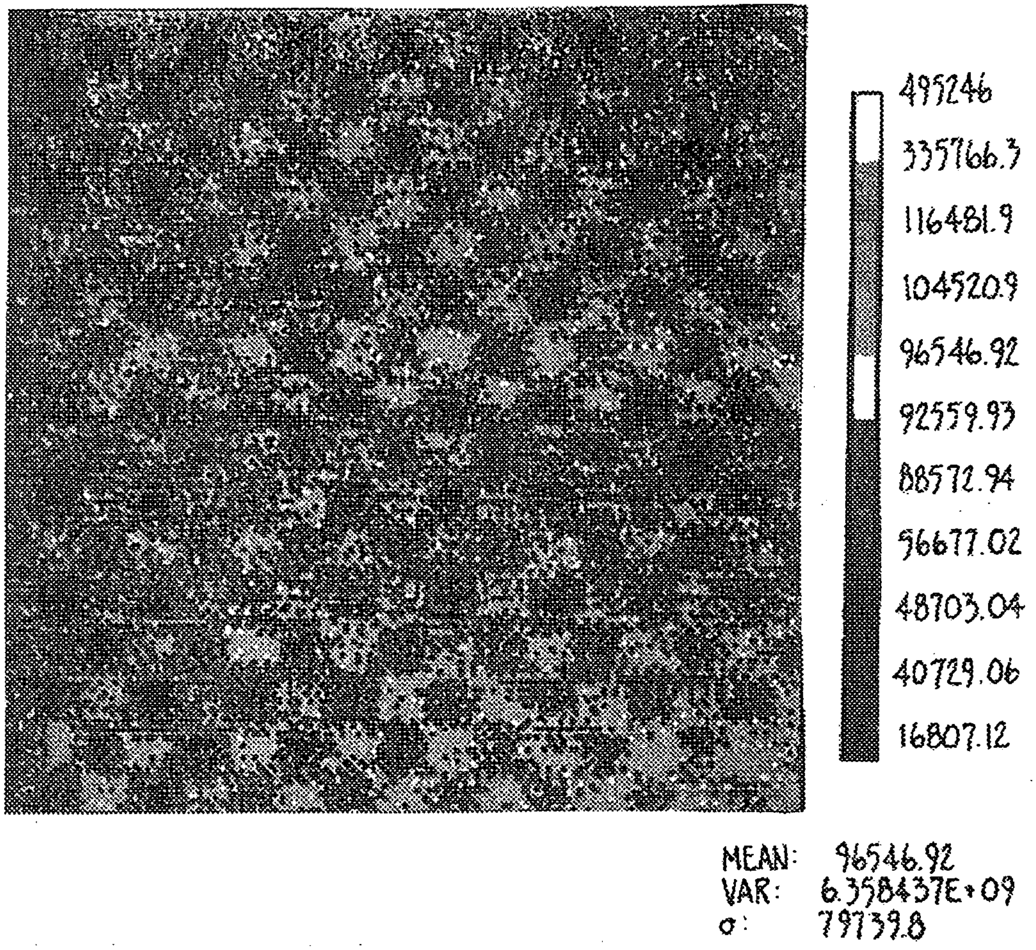

495246
335766.3
116481.9
104520.9
96546.92
92559.93
88572.94
56677.02
48703.04
40729.06
16807.12
MEAN: 96546.92
VAR: 6.358437E+09
σ: 79739.8

FIG._13D.

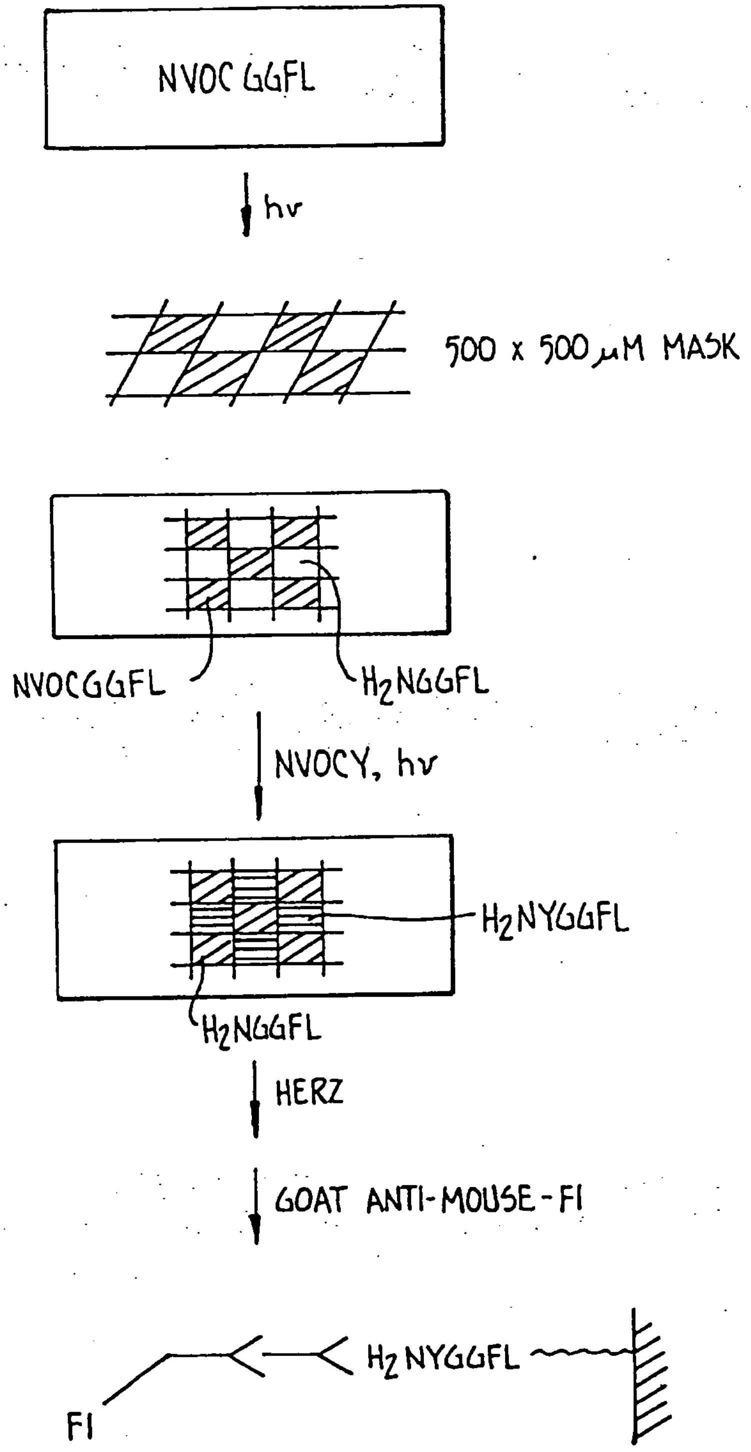
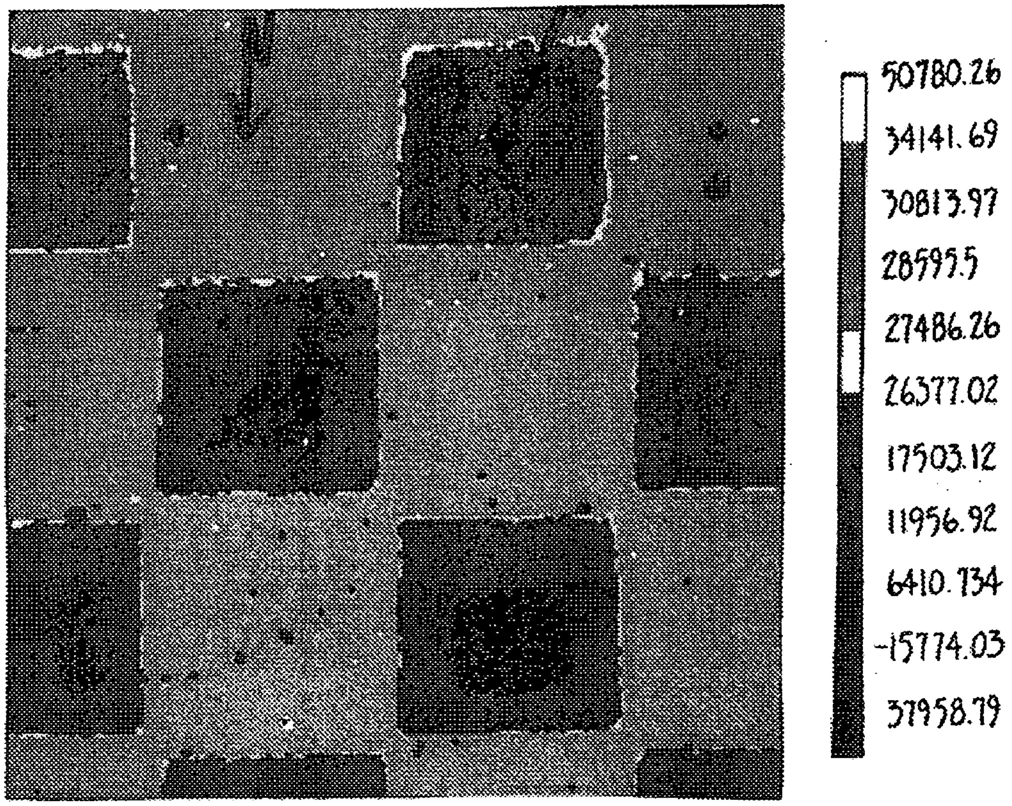
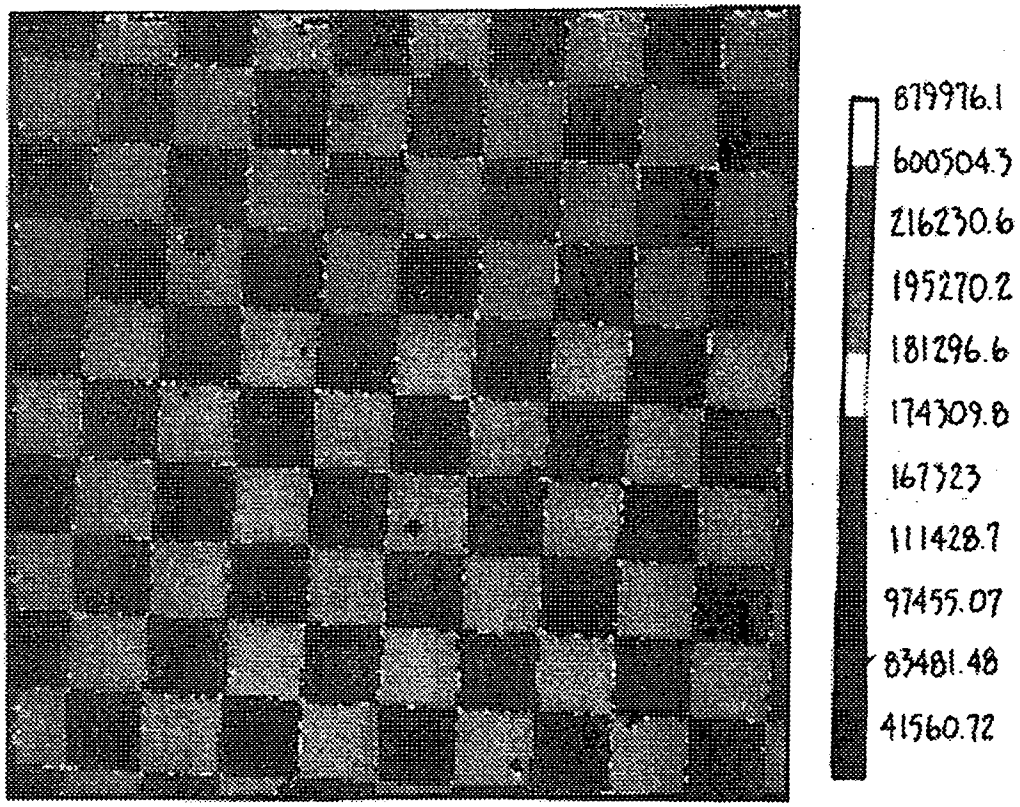

NVOC GGFL
hv
500 x 500 μM MASK
NVOCGGFL
H2NGGFL
NVOCY, hv
H2NYGGFL
H2NGGFL
HERZ
GOAT ANTI-MOUSE-FI
H2NYGGFL
FI

FIG._14B.



MEAN: 28595.5
VAR: 4.921637E+08
σ: 22184.76

FIG_15A.



MEAN: 181296.6
VAR: 1.952612E+10
σ: 139735.9

FIG._15B.

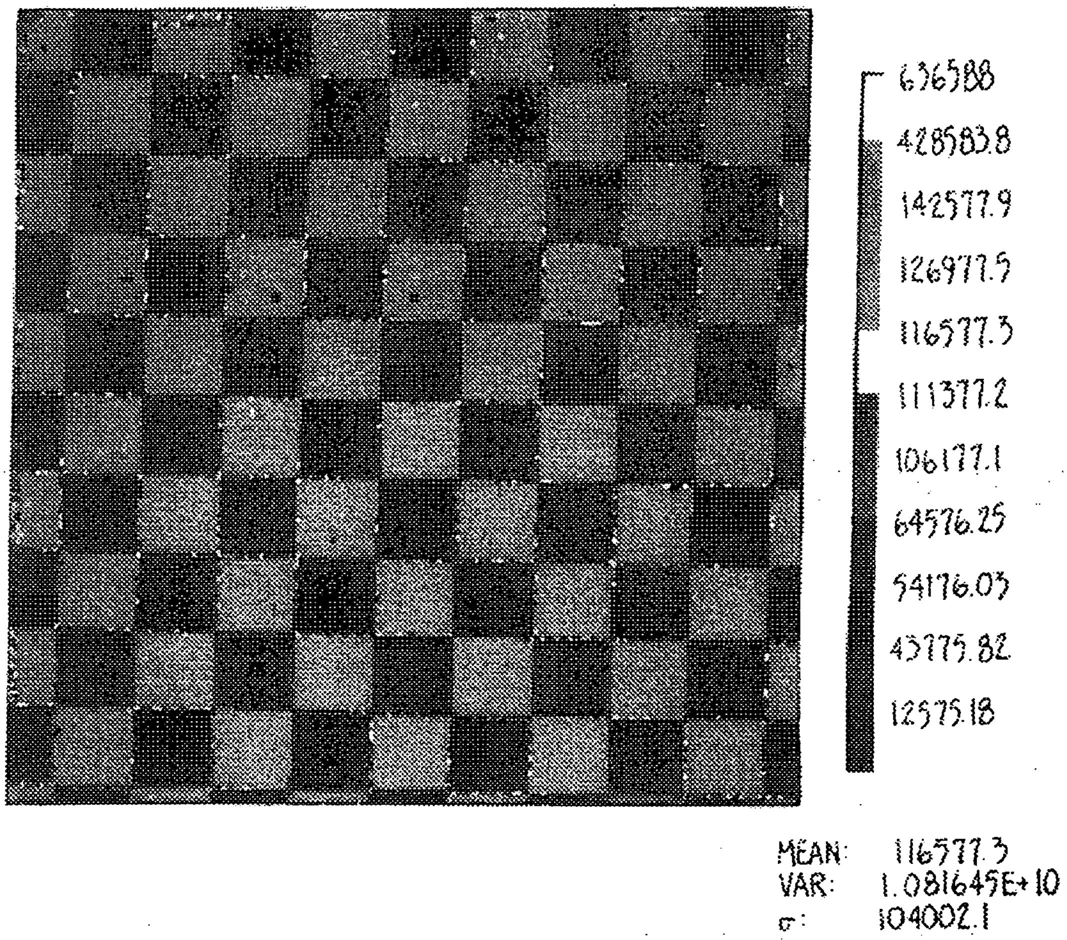
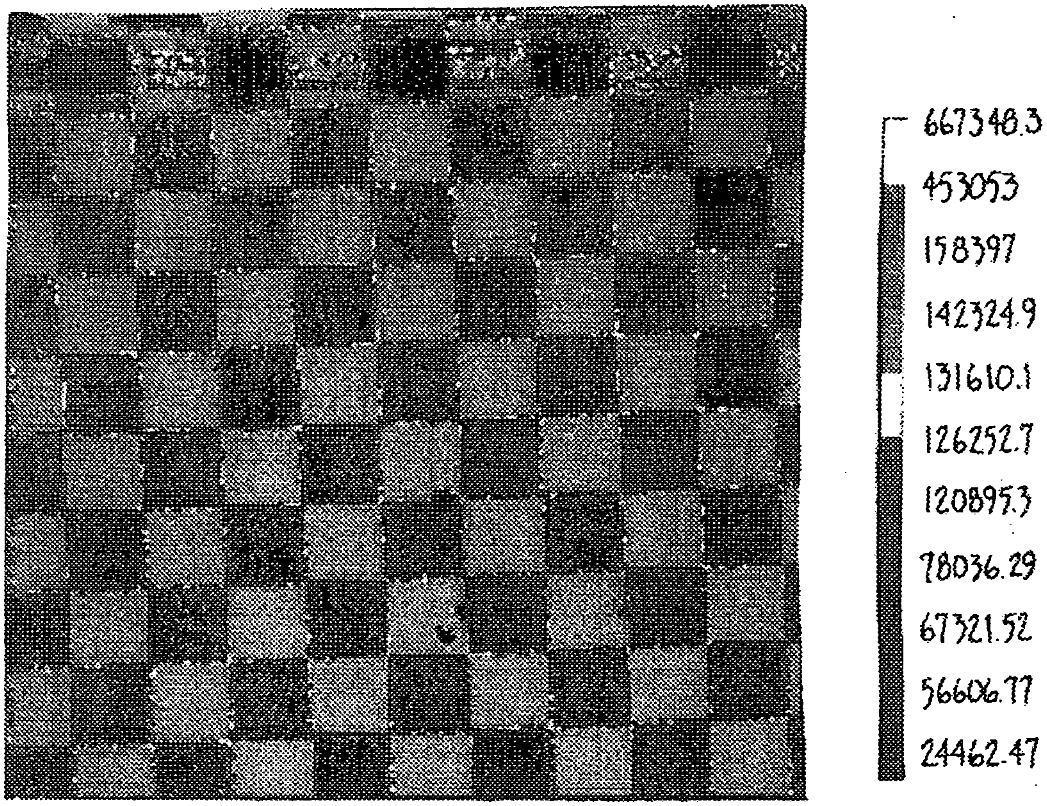

636588
428583.8
142577.9
126977.5
116577.3
111377.2
106177.1
64576.25
54176.03
43775.82
12575.18
MEAN: 116577.3
VAR: 1.081645E+10
σ: 104002.1

FIG._16.



MEAN: 131610.1
VAR: 1.148062E+10
σ: 107147.6

FIG._17.

P	A	S	G	
LPGFL	LAGFL	LSGFL	LGGFL	L
FPGFL	FAGFL	FSGFL	FGGFL	F
WPGFL	WAGFL	WSGFL	WGGFL	W
YPGFL	YAGFL	YSGFL	YGGFL	Y

L SET

FIG._18A.

p	a	s	G	
YpGFL	YaGFL	YsGFL	YGGFL	Y
fpGFL	faGFL	fsGFL	fGGFL	f
wpGFL	waGFL	wsGFL	wGGFL	w
ypGFL	yaGFL	ysGFL	yGGFL	y

D SET

FIG._18B.

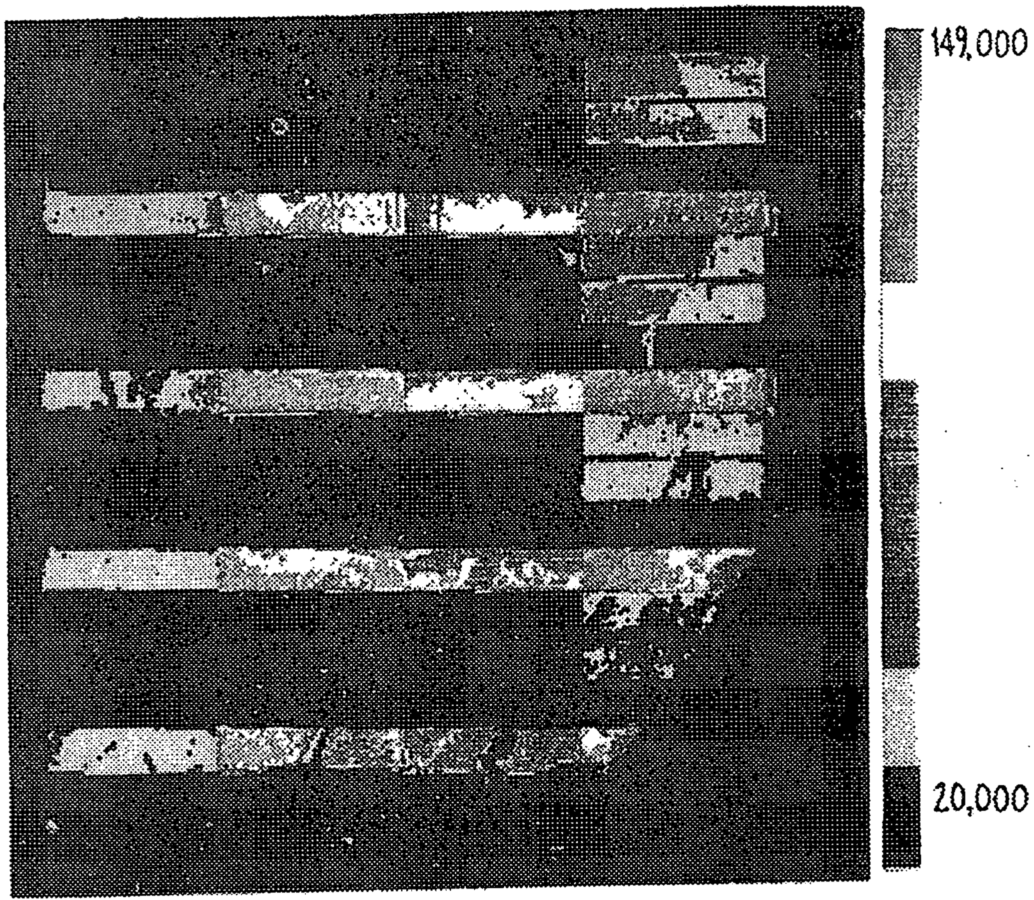


FIG._19.

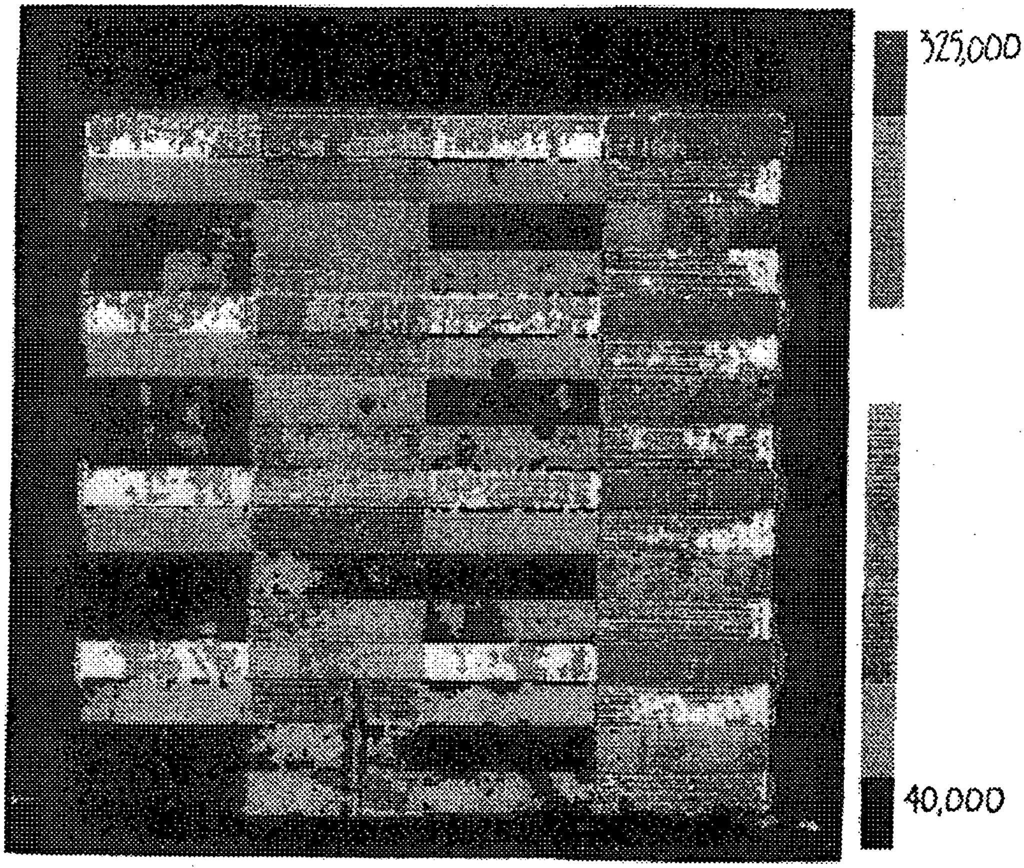


FIG._20.

NUCLEIC ACID ARRAYS

This application is a continuation of Ser. No. 09/129,470 filed Aug. 4, 1998, which is a continuation of Ser. No. 08/456,598 filed Jun. 1, 1995, which is a divisional of Ser. No. 07/954,646 filed Sep. 30, 1992, now issued as U.S. Pat. No. 5,445,934, which is a divisional of Ser. No. 07/850,356, filed Mar. 12, 1992, now issued as U.S. Pat. No. 5,405,783, which is a divisional of Ser. No. 07/492,462 filed Mar. 7, 1990, now issued as U.S. Pat. No. 5,143,854, which is a continuation-in-part of Ser. No. 07/362,901 filed Jun. 7, 1989, now abandoned, the disclosures of which are incorporated by reference.

COPYRIGHT NOTICE

BACKGROUND OF THE INVENTION

The present inventions relate to the synthesis and placement materials at known locations. In particular, one embodiment of the inventions provides a method and associated apparatus for preparing diverse chemical sequences at known locations on a single substrate surface. The inventions may be applied, for example, in the field of preparation of oligomer, peptide, nucleic acid, oligosaccharide, phospholipid, polymer, or drug congener preparation, especially to create sources of chemical diversity for use in screening for biological activity.

The relationship between structure and activity of molecules is a fundamental issue in the study of biological systems. Structure-activity relationships are important in understanding, for example, the function of enzymes, the ways in which cells communicate with each other, as well as cellular control and feedback systems.

Certain macromolecules are known to interact and bind to other molecules having a very specific three-dimensional spatial and electronic distribution. Any large molecule having such specificity can be considered a receptor, whether it is an enzyme catalyzing hydrolysis of a metabolic intermediate, a cell-surface protein mediating membrane transport of ions, a glycoprotein serving to identify a particular cell to its neighbors, an IgG-class antibody circulating in the plasma, an oligonucleotide sequence of DNA in the nucleus, or the like. The various molecules which receptors selectively bind are known as ligands.

Many assays are available for measuring the binding affinity of known receptors and ligands, but the information which can be gained from such experiments is often limited by the number and type of ligands which are available. Novel ligands are sometimes discovered by chance or by application of new techniques for the elucidation of molecular structure, including x-ray crystallographic analysis and recombinant genetic techniques for proteins.

Small peptides are an exemplary system for exploring the relationship between structure and function in biology. A peptide is a sequence of amino acids. When the twenty naturally occurring amino acids are condensed into polymeric molecules they form a wide variety of three-dimensional configurations, each resulting from a particular amino acid sequence and solvent condition. The number of possible pentapeptides of the 20 naturally occurring amino acids, for example, is 20^5 or 3.2 million different peptides. The likelihood that molecules of this size might be useful in receptor-binding studies is supported by epitope analysis studies showing that some antibodies recognize sequences as short as a few amino acids with high specificity. Furthermore, the average molecular weight of amino acids puts small peptides in the size range of many currently useful pharmaceutical products.

Pharmaceutical drug discovery is one type of research which relies on such a study of structure-activity relationships. In most cases, contemporary pharmaceutical research can be described as the process of discovering novel ligands with desirable patterns of specificity for biologically important receptors. Another example is research to discover new compounds for use in agriculture, such as pesticides and herbicides.

Sometimes, the solution to a rational process of designing ligands is difficult or unyielding. Prior methods of preparing large numbers of different polymers have been painstakingly slow when used at a scale sufficient to permit effective rational or random screening. For example, the "Merrifield" method (*J. Am. Chem. Soc.* (1963) 85:2149–2154, which is incorporated herein by reference for all purposes) has been used to synthesize peptides on a solid support. In the Merrifield method, an amino acid is covalently bonded to a support made of an insoluble polymer. Another amino acid with an alpha protected group is reacted with the covalently bonded amino acid to form a dipeptide. After washing, the protective group is removed and a third amino acid with an alpha protective group is added to the dipeptide. This process is continued until a peptide of a desired length and sequence is obtained. Using the Merrifield method, it is not economically practical to synthesize more than a handful of peptide sequences in a day.

To synthesize larger numbers of polymer sequences, it has also been proposed to use a series of reaction vessels for polymer synthesis. For example, a tubular reactor system may be used to synthesize a linear polymer on a solid phase support by automated sequential addition of reagents. This method still does not enable the synthesis of a sufficiently large number of polymer sequences for effective economical screening.

Methods of preparing a plurality of polymer sequences are also known in which a foraminous container encloses a known quantity of reactive particles, the particles being larger in size than foramina of the container. The containers may be selectively reacted with desired materials to synthesize desired sequences of product molecules. As with other methods known in the art, this method cannot practically be used to synthesize a sufficient variety of polypeptides for effective screening.

Other techniques have also been described. These methods include the synthesis of peptides on 96 plastic pins which fit the format of standard microtiter plates. Unfortunately, while these techniques have been somewhat useful, substantial problems remain. For example, these methods continue to be limited in the diversity of sequences which-can be economically synthesized and screened.

From the above, it is seen that an improved method and apparatus for synthesizing a variety of chemical sequences at known locations is desired.

SUMMARY OF THE INVENTION

An improved method and apparatus for the preparation of a variety of polymers is disclosed.

In one preferred embodiment, linker molecules are provided on a substrate. A terminal end of the linker molecules is provided with a reactive functional group protected with a photoremovable protective group. Using lithographic methods, the photoremovable protective group is exposed to light and removed from the linker molecules in first selected regions. The substrate is then washed or otherwise contacted with a first monomer that reacts with exposed functional groups on the linker molecules. In a preferred embodiment, the monomer is an amino acid containing a photoremovable protective group at its amino or carboxy terminus and the linker molecule terminates in an amino or carboxy acid group bearing a photoremovable protective group.

A second set of selected regions is, thereafter, exposed to light and the photoremovable protective group on the linker molecule/protected amino acid is removed at the second set of regions. The substrate is then contacted with a second monomer containing a photoremovable protective group for reaction with exposed functional groups. This process is repeated to selectively apply monomers until polymers of a desired length and desired chemical sequence are obtained. Photolabile groups are then optionally removed and the sequence is, thereafter, optionally capped. Side chain protective groups, if present, are also removed.

By using the lithographic techniques disclosed herein, it is possible to direct light to relatively small and precisely known locations on the substrate. It is, therefore, possible to synthesize polymers of a known chemical sequence at known locations on the substrate.

The resulting substrate will have a variety of uses including, for example, screening large numbers of polymers for biological activity. To screen for biological activity, the substrate is exposed to one or more receptors such as antibody whole cells, receptors on vesicles, lipids, or any one of a variety of other receptors. The receptors are preferably labeled with, for example, a fluorescent marker, radioactive marker, or a labeled antibody reactive with the receptor. The location of the marker on the substrate is detected with, for example, photon detection or autoradiographic techniques. Through knowledge of the sequence of the material at the location where binding is detected, it is possible to quickly determine which sequence binds with the receptor and, therefore, the technique can be used to screen large numbers of peptides. Other possible applications of the inventions herein include diagnostics in which various antibodies for particular receptors would be placed on a substrate and, for example, blood sera would be screened for immune deficiencies. Still further applications include, for example, selective "doping" of organic materials in semiconductor devices, and the like.

In connection with one aspect of the invention an improved reactor system for synthesizing polymers is also disclosed. The reactor system includes a substrate mount which engages a substrate around a periphery thereof. The substrate mount provides for a reactor space between the substrate and the mount through or into which reaction fluids are pumped or flowed. A mask is placed on or focused on the substrate and illuminated so as to deprotect selected regions of the substrate in the reactor space. A monomer is pumped through the reactor space or otherwise contacted with the substrate and reacts with the deprotected regions. By selectively deprotecting regions on the substrate and flowing predetermined monomers through the reactor space, desired polymers at known locations may be synthesized.

Improved detection apparatus and methods are also disclosed. The detection method and apparatus utilize a substrate having a large variety of polymer sequences at known locations on a surface thereof. The substrate is exposed to a fluorescently labeled receptor which binds to one or more of the polymer sequences. The substrate is placed in a microscope detection apparatus for identification of locations where binding takes place. The microscope detection apparatus includes a monochromatic or polychromatic light source for directing light at the substrate, means for detecting fluoresced light from the substrate, and means for determining a location of the fluoresced light. The means for detecting light fluoresced on the substrate may in some embodiments include a photon counter. The means for determining a location of the fluoresced light may include an x/y translation table for the substrate. Translation of the slide and data collection are recorded and managed by an appropriately programmed digital computer.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1 illustrates masking and irradiation of a substrate at a first location. The substrate is shown in cross-section;

FIG. 2 illustrates the substrate after application of a monomer "A";

FIG. 3 illustrates irradiation of the substrate at a second location;

FIG. 4 illustrates the substrate after application of monomer "B";

FIG. 5 illustrates irradiation of the "A" monomer;

FIG. 6 illustrates the substrate after a second application of "B";

FIG. 7 illustrates a completed substrate;

FIGS. 8A and 8B illustrate alternative embodiments of a reactor system for forming a plurality of polymers on a substrate;

FIG. 9 illustrates a detection apparatus for locating fluorescent markers on the substrate;

FIGS. 10A–10M illustrate the method as it is applied to the production of the trimers of monomers "A" and "B";

FIGS. 11A and 11B are fluorescence traces for standard fluorescent beads;

FIGS. 12A and 12B are fluorescence curves for NVOC slides not exposed and exposed to light respectively;

FIGS. 13A to 13D are fluorescence plots of slides exposed through 100 μm, 50 μm, 20 μm, and 10 μm masks;

FIG. 14A and 14B illustrates fluorescence of a slide pith the peptide YGGFL on selected regions of-its surface which has been exposed to labeled Herz antibody specific for this sequence;

FIGS. 15A and 15D illustrate formation of and a fluorescence plot of a slide with a checkerboard pattern of YGGFL and GGFL exposed to labeled Herz antibody. FIG. 15A illustrates a 500×500 μm mask which has been focused on the substrate according to FIG. 8A while FIG. 15B illustrates a 50×50 μm mask placed in direct contact with the substrate in accord with FIG. 8B;

FIG. 16 is a fluorescence plot of YGGFL and PGGFL synthesized in a 50 μm checkerboard pattern;

FIG. 17 is a fluorescence plot of YPGGFL and is YGGFL synthesized in a 50 μm checkerboard pattern;

FIGS. 18A and 18B illustrate the mapping of sixteen sequences synthesized on two different glass slides;

FIG. 19 is a fluorescence plot of the slide illustrated in FIG. 18A; and

FIG. 20 is a fluorescence plot of the slide illustrated in FIG. 10B.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

CONTENTS

I. Glossary

The following terms are intended to have the following general meanings as they are used herein:

1. Complementary: Refers to the topological compatibility or matching together of interacting surfaces of a ligand molecule and its receptor. Thus, the receptor and its ligand can be described as complementary, and furthermore, the contact surface characteristics are complementary to each other.
2. Epitope: The portion of an antigen molecule which is delineated by the area of interaction with the subclass of receptors known as antibodies.
3. Ligand: A ligand is a molecule that is recognized by a particular receptor. Examples of ligands that can be investigated by this invention include, but are not restricted to, agonists and antagonists for cell membrane receptors, toxins and venoms, viral epitopes, hormones (e.g., opiates, steroids, etc.), hormone receptors, peptides, enzymes, enzyme substrates, cofactors, drugs, lectins, sugars, oligonucleotides, nucleic acids, oligosaccharides, proteins, and monoclonal antibodies.
4. Monomer: A member of the set of small molecules which can be joined together to form a polymer. The set of monomers includes but is not restricted to, for example, the set of common L-amino acids, the set of D-amino acids, the set of synthetic amino acids, the set of nucleotides and the set of pentoses and hexoses. As used herein, monomers refers to any member of a basis set for synthesis of a polymer. For example, dimers of L-amino acids form a basis set of 400 monomers for synthesis of polypeptides. Different basis sets of monomers may be used at successive steps in the synthesis of a polymer.
5. Peptide: A polymer in which the monomers are alpha amino acids and which are joined together through amide bonds and alternatively referred to as a polypeptide. In the context of this specification it should be appreciated that the amino acids may be the L-optical isomer or the D-optical isomer. Peptides are more than two amino acid monomers long, and often more than 20 amino acid monomers long. Standard abbreviations for amino acids are used (e.g., P for proline). These abbreviations are included in Stryer, *Biochemstry*, Third Ed., 1988, which is incorporated herein by reference for all purposes.
6. Radiation: Energy which may be selectively applied including energy having a wavelength of between 10^{-14} and 10^{4} meters including, for example, electron beam radiation, gamma radiation, x-ray radiation, ultra-violet radiation, visible light, infrared radiation, microwave radiation, and radio waves. "Irradiation" refers to the application of radiation to a surface.
7. Receptor: A molecule that has an affinity for a given ligand. Receptors may be naturally-occuring or manmade molecules. Also, they can be employed in their unaltered state or as aggregates with other species. Receptors may be attached, covalently or noncovalently, to a binding member, either directly or via a specific binding substance. Examples of receptors which can be employed by this invention include, but are not restricted to, antibodies, cell membrane receptors, monoclonal antibodies and antisera reactive with specific antigenic determinants (such as on viruses, cells or other materials), drugs, polynucleotides, nucleic acids, peptides, cofactors, lectins, sugars, polysaccharides, cells, cellular membranes, and organelles. Receptors are sometimes referred to in the art as anti-ligands. As the term receptors is used herein, no difference in meaning is intended. A "Ligand Receptor Pair" is formed when two macromolecules have combined through molecular recognition to form a complex.

Other examples of receptors which can be investigated by this invention include but are not restricted to:

a) Microorganism receptors: Determination of ligands which bind to receptors, such as specific transport proteins or enzymes essential to survival of microorganisms, is useful in a new class of antibiotics. Of particular value would be antibiotics against opportunistic fungi, protozoa, and those bacteria resistant to the antibiotics in current use.

b) Enzymes: For instance, the binding site of enzymes such as the enzymes responsible for cleaving neurotransmitters; determination of ligands which bind to certain receptors to modulate the action of the enzymes which cleave the different neurotransmitters is useful in the development of drugs which can be used in the treatment of disorders of neurotransmission.

c) Antibodies: For instance, the invention may be useful in investigating the ligand-binding site on the antibody molecule which combines with the epitope of an antigen of interest; determining a sequence that mimics an antigenic epitope may lead to the development of vaccines of which the immunogen is based on one or more of such sequences or lead to the development of related diagnostic agents or compounds useful in therapeutic treatments such as for auto-immune diseases (e.g., by blocking the binding of the "self" antibodies).

d) Nucleic Acids: Sequences of nucleic acids may be synthesized to establish DNA or RNA binding sequences.

e) Catalytic Polypeptides: Polymers, preferably polypeptides, which are capable of promoting a chemical reaction involving the conversion of one or more reactants to one or more products. Such polypeptides generally include a binding site specific for at least one reactant or reaction intermediate and an active functionality proximate to the binding site, which functionality is capable of chemically modifying the bound reactant. Catalytic polypeptides are described in, for example, U.S. application Ser. No. 404,920, which is incorporated herein by reference for all purposes.

f) Hormone receptors: For instance, the receptors for insulin and growth hormone. Determination of the ligands which bind with high affinity to a receptor is useful in the development of, for example, an oral replacement of the daily injections which diabetics must take to relieve the symptoms of diabetes, and in the other case, a replacement for the scarce human growth hormone which can only be obtained from cadavers or by recombinant DNA technology. Other examples are the vasoconstrictive hormone receptors; determination of those ligands which bind to a receptor may lead to the development of drugs to control blood pressure.

g) Opiate receptors: Determination of ligands which bind to the opiate receptors in the brain is useful in the development of less-addictive replacements for morphine and related drugs.

8. Substrate: A material having a rigid or semi-rigid surface. In many embodiments, at least one surface of the substrate will be substantially flat, although in some embodiments it may be desirable to physically separate synthesis regions for different polymers with, for example, wells, raised regions, etched trenches, or the like. According to other embodiments, small beads may be provided on the surface which may be released upon completion of the synthesis.

9. Protective Group: A material which is bound to a monomer unit and which may be spatially removed upon selective exposure to an activator such as electromagnetic radiation. Examples of protective groups with utility herein include Nitroveratryloxy carbonyl, Nitrobenzyloxy carbonyl, Dimethyl dimethoxybenzyloxy carbonyl, 5-Bromo-7-nitroindolinyl, o-Hydroxy-α-methyl cinnamoyl, and 2-oxymethylene anthraquinone. Other examples of activators include ion beams, electric fields, magnetic fields, electron beams, x-ray, and the like.

10. Predefined Region: A predefined region is a localized area on a surface which is, was, or is intended to be activated for formation of a polymer. The predefined region may have any convenient shape, e.g., circular, rectangular, elliptical, wedge-shaped, etc. For the sake of brevity herein, "predefined regions" are sometimes referred to simply as "regions."

11. Substantially Pure: A polymer is considered to be "substantially pure" within a predefined region of a substrate when it exhibits characteristics that distinguish it from other predefined regions. Typically, purity will be measured in terms of biological activity or function as a result of uniform sequence. Such characteristics will typically be measured by way of binding with a selected ligand or receptor.

II. General

The present invention provides methods and apparatus for the preparation and use of a substrate having a plurality of polymer sequences in predefined regions. The invention is described herein primarily with regard to the preparation of molecules containing sequences of amino acids, but could readily be applied in the preparation of other polymers. Such polymers include, for example, both linear and cyclic polymers of nucleic acids, polysaccharides, phospholipids, and peptides having either α-, β-, or ω-amino acids, heteropolymers in which a known drug is covalently bound to any of the above, polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneimines, polyarylene sulfides, polysiloxanes, polyimides, polyacetates, or other polymers which will be apparent upon review of this disclosure. In a preferred embodiment, the invention herein is used in the synthesis of peptides.

The prepared substrate may, for example, be used in screening a variety of polymers as ligands for binding with a receptor, although it will be apparent that the invention could be used for the synthesis of a receptor for binding with a ligand. The substrate disclosed herein will have a wide variety of other uses. Merely by way of example, the invention herein can be used in determining peptide and nucleic acid sequences which bind to proteins, finding sequence-specific binding drugs, identifying epitopes recognized by antibodies, and evaluation of a variety of drugs for clinical and diagnostic applications, as well as combinations of the above.

The invention preferably provides for the use of a substrate "S" with a surface. Linker molecules "L" are optionally provided on a surface of the substrate. The purpose of the linker molecules, in some embodiments, is to facilitate receptor recognition of the synthesized polymers.

Optionally, the linker molecules may be chemically protected for storage purposes. A chemical storage protective group such as t-BOC (t-butoxycarbonyl) may be used in some embodiments. Such chemical protective groups would be chemically removed upon exposure to, for example, acidic solution and would serve to protect the surface during storage and be removed prior to polymer preparation.

On the substrate or a distal end of the linker molecules, a functional group with a protective group P_0 is provided. The protective group P_0 may be removed upon exposure to radiation, electric fields, electric currents, or other activators to expose the functional group.

In a preferred embodiment, the radiation is ultraviolet (UV), infrared (IR), or visible light. As more fully described below, the protective group may alternatively be an electrochemically-sensitive group which may be removed in the presence of an electric field. In still further alternative embodiments, ion beams, electron beams, or the like may be used for deprotection.

In some embodiments, the exposed regions and, therefore, the area upon which each distinct polymer sequence is synthesized are smaller than about 1 cm^2 or less than 1 mm^2. In preferred embodiments the exposed area is less than about 10,000 μm^2 or, more preferably, less than 100 μm^2 and may, in some embodiments, encompass the binding site for as few as a single molecule. Within these regions, each polymer is preferably synthesized in a substantially pure form.

Concurrently or after exposure of a known region of the substrate to light, the surface is contacted with a first monomer unit M_1 which reacts with the functional group which has been exposed by the deprotection step. The first monomer includes a protective group P_1. P_1 may or may not be the same as P_0.

Accordingly, after a first cycle, known first regions of the surface may comprise the sequence:

$S\text{-}L\text{-}M_1\text{-}P_1$

while remaining regions of the surface comprise the sequence:

$S\text{-}L\text{-}P_0$.

Thereafter, second regions of the surface (which may include the first region) are exposed to light and contacted with a second monomer M_2 (which may or may not be the same as M_1) having a protective group P_2. P_2 may or may not be the same as P_0 and P_1. After this second cycle, different regions of the substrate may comprise one or more of the following sequences:

$S\text{-}L\text{-}M_1\text{-}M_2\text{-}P_2$

$S\text{-}L\text{-}M_2\text{-}P_2$

$S\text{-}L\text{-}M_1\text{-}P_1$ and/or

$S\text{-}L\text{-}P_0$.

The above process is repeated until the substrate includes desired polymers of desired lengths. By controlling the locations of the substrate exposed to light and the reagents exposed to the substrate following exposure, the location of each sequence will be known.

Thereafter, the protective groups are removed from some or all of the substrate and the sequences are, optionally, capped with a capping unit C. The process results in a substrate having a surface with a plurality of polymers of the following general formula:

$S\text{-}[L]\text{-}(M_i)\text{-}(M_j)\text{-}(M_k) \ldots (M_x)\text{-}[C]$

where square brackets indicate optional groups, and $M_i \ldots M_x$ indicates any sequence of monomers. The number of monomers could cover a wide variety of values, but in a preferred embodiment they will range from 2 to 100.

In some embodiments a plurality of locations on the substrate polymers are to contain a common monomer subsequence. For example, it may be desired to synthesize a sequence $S\text{-}M_1\text{-}M_2\text{-}M_3$ at first locations and a sequence $S\text{-}M_4\text{-}M_2\text{-}M_3$ at second locations. The process would commence with irradiation of the first locations followed by contacting with M_1-P, resulting in the sequence $S\text{-}M_1$-P at the first location. The second locations would then be irradiated and contacted with M_4-P, resulting in the sequence $S\text{-}M_4$-P at the second locations. Thereafter both the first and second locations would be irradiated and contacted with the dimer $M_2\text{-}M_3$, resulting in the sequence $S\text{-}M_1\text{-}M_2\text{-}M_3$ at the first locations and $S\text{-}M_4\text{-}M_2\text{-}M_3$ at the second locations. Of course, common subsequences of any length could be utilized including those in a range of 2 or more monomers, 2 to 100 monomers, 2 to 20 monomers, and a most preferred range of 2 to 3 monomers.

According to other embodiments, a set of masks is used for the first monomer layer and, thereafter, varied light wavelengths are used for selective deprotection. For example, in the process discussed above, first regions are first exposed through a mask and reacted with a first monomer having a first protective group P_1, which is removable upon exposure to a first wavelength of light (e.g., IR). Second regions are masked and reacted with a second monomer having a second protecive group P_2, which is removable upon exposure to a second wavelength of light (e.g., UV). Thereafter, masks become unnecessary in the synthesis because the entire substrate may be exposed alternatively to the first and second wavelengths of light in the deprotection cycle.

The polymers prepared on a substrate according to the above methods will have a variety of uses including, for example, screening for biological activity. In such screening activities, the substrate containing the sequences is exposed to an unlabeled or labeled receptor such as an antibody, receptor on a cell, phospholipid vesicle, or any one of a variety of other receptors. In one preferred embodiment the polymers are exposed to a first, unlabeled receptor of interest and, thereafter, exposed to a labeled receptor-specific recognition element, which is, for example, an antibody. This process will provide signal amplification in the detection stage.

The receptor molecules may bind with one or more polymers on the substrate. The presence of the labeled receptor and, therefore, the presence of a sequence which binds with the receptor is detected in a preferred embodiment through the use of autoradiography, detection of fluorescence with a charge-coupled device, fluorescence microscopy, or the like. The sequence of the polymer at the locations where the receptor binding is detected may be used to determine all or part of a sequence which is complementary to the receptor.

Use of the invention herein is illustrated primarily with reference to screening for biological activity. The invention will, however, find many other uses. For example, the invention may be used in information storage (e.g., on optical disks), production of molecular electronic devices, production of stationary phases in separation sciences, production of dyes and brightening agents, photography, and in immobilization of cells, proteins, lectins, nucleic acids, polysaccharides and the like in patterns on a surface via molecular recognition of specific polymer sequences. By synthesizing the same compound in adjacent, progressively differing concentrations, a gradient will be established to control chemotaxis or to develop diagnostic dipsticks which, for example, titrate an antibody against an increasing amount of antigen. By synthesizing several catalyst molecules in close proximity, more efficient multistep conversions may be achieved by "coordinate immobilization." Coordinate immobilization also may be used for electron transfer systems, as well as to provide both structural integrity and other desirable properties to materials such as lubrication, wetting, etc.

According to alternative embodiments, molecular biodistribution or pharmacokinetic properties may be examined. For example, to assess resistance to intestinal or serum proteases, polymers may be capped with a fluorescent tag and exposed to biological fluids of interest.

III. Polymer Synthesis

FIG. 1 illustrates one embodiment of the invention disclosed herein in which a substrate 2 is shown in cross-section. Essentially, any conceivable substrate may be employed in the invention. The substrate may be biological, nonbiological, organic, inorganic, or a combination of any of these, existing as particles, strands, precipitates, gels, sheets, tubing, spheres, containers, capillaries, pads, slices, films, plates, slides, etc. The substrate may have any convenient shape, such as a disc, square, sphere, circle, etc. The substrate is preferably flat but may take on a variety of alternative surface configurations. For example, the substrate may contain raised or depressed regions on which the synthesis takes place. The substrate and its surface preferably form a rigid support on which to carry out the reactions described herein. The substrate and its surface is also chosen to provide appropriate light-absorbing characteristics. For instance, the substrate may be a polymerized Langmuir Blodgett film, functionalized glass, Si, Ge, GaAs, GaP, SiO_2, SiN_4, modified silicon, or any one of a wide variety of gels or polymers such as (poly)tetrafluoroethylene, (poly) vinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof. Other substrate materials will be readily apparent to those of skill in is the art upon review of this disclosure. In a preferred embodiment the substrate is flat glass or single-crystal silicon with surface relief features of less than 10 Å.

According to some embodiments, the surface of the substrate is etched using well known techniques to provide for desired surface features. For example, by way of the formation of trenches, v-grooves, mesa structures, or the like, the synthesis regions may be more closely placed within the focus point of impinging light, be provided with reflective "mirror" structures for maximization of light collection from fluorescent sources, or the like.

Surfaces on the solid substrate will usually, though not always, be composed of the same material as the substrate. Thus, the surface may be composed of any of a wide variety of materials, for example, polymers, plastics, resins, polysaccharides, silica or silica-based materials, carbon, metals, inorganic glasses, membranes, or any of the above-listed substrate materials. In some embodiments the surface may provide for the use of caged binding members which are attached firmly to the surface of the substrate in accord with the teaching of copending application Ser. No. 404,920, previously incorporated herein by reference. Preferably, the surface will contain reactive groups, which could be carboxyl, amino, hydroxyl, or the like. Most preferably, the surface will be optically transparent and will have surface Si—OH functionalities, such as are found on silica surfaces.

The surface 4 of the substrate is preferably provided with a layer of linker molecules 6, although it will be understood that the linker molecules are not required elements of the invention. The linker molecules are preferably of sufficientlength to permit polymers in a completed substrate to interact freely with molecules exposed to the substrate. The linker molecules should be 6–50 atoms long to provide sufficient exposure. The linker molecules may be, for example, aryl acetylene, ethylene glycol oligomers containing 2–10 monomer units, diamines, diacids, amino acids, or combinations thereof. Other linker molecules may be used in light of this disclsoure.

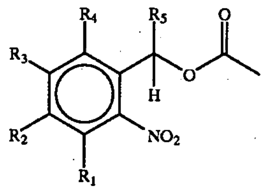
According to alternative embodiments, the linker molecules are selected based upon their hydrophilic/hydrophobic properties to improve presentation of synthesized polymers to certain receptors. For example, in the case of a hydrophilic receptor, hydrophilic linker molecules will be preferred so as to permit the receptor to more closely approach the synthesized polymer.

According to another alternative embodiment, linker molecules are also provided with a photocleavable group at an intermediate position. The photocleavable group is preferably cleavable at a wavelength different from the protective group. This enables removal of the various polymers following completion of the synthesis by way of exposure to the different wavelengths of light.

The linker molecules can be attached to the substrate via carbon-carbon bonds using, for example, (poly) trifluorochloroethylene surfaces, or preferably, by siloxane bonds (using, for example, glass or silicon oxide surfaces). Siloxane bonds with the surface of the substrate may be formed in one embodiment via reactions of linker molecules bearing trichlorosilyl groups. The linker molecules may optionally be attached in an ordered array, i.e., as parts of the head groups in a polymerized Langmuir Blodgett film. In alternative embodiments, the linker molecules are adsorbed to the surface of the substrate.

The linker molecules and monomers used herein are provided with a functionial group to which is bound a protective group. Preferably, the protective group is on the distal or terminal end of the linker molecule opposite the substrate. The protective group may be either a negative protective group (i.e., the protective group renders the linker molecules less reactive with a monomer upon exposure) or a positive protective group (i.e., the protective group renders the linker molecules more reactive with a monomer upon exposure). In the case of negative protective groups an additional step of reactivation will be required. In some embodiments, this will be done by heating.

The protective group on the linker molecules may be selected from a wide variety of positive light-reactive groups preferably including nitro aromatic compounds such as o-nitrobenzyl derivatives or benzylsulfonyl. In a preferred embodiment, 6-nitroveratryloxy-carbonyl (NVOC), 2-nitrobenzyloxycarbonyl (NBOC) or α,α-dimethyl-dimethoxybenzyloxycarbonyl (DDZ) is used. In one embodiment, a nitro aromatic compound containing a benzylic hydrogen ortho to the nitro group is used, i.e., a chemical of the form:



where R_1 is alkoxy, alkyl, halo, aryl, alkenyl, or hydrogen; R_2 is alkoxy, alkyl, halo, aryl, nitro, or hydrogen; R_3 is alkoxy, alkyl, halo, nitro, aryl, or hydrogen; R_4 is alkoxy, alkyl, hydrogen, aryl, halo, or nitro; and R_5 is alkyl, alkynyl, cyano, alkoxy, hydrogen, halo, aryl, or alkenyl. Other materials which may be used include o-hydroxy-α-methyl cinnamoyl derivatives. Photoremovable protective groups are described in, for example, Patchornik, *J. Am. Chem. Soc.* (1970) 92:6333 and Amit et al., *J. Org. Chem.* (1974) 39:192, both of which are incorporated herein by reference.

In an alternative embodiment the positive reactive group is activated for reaction with reagents in solution. For example, a 5-bromo-7-nitro indoline group, when bound to a carbonyl, undergoes reaction upon exposure to light at 420 nm.

In a second alternative embodiment, the reactive group on the linker molecule is selected from a wide variety of negative light-reactive groups including a cinammate group.

Alternatively, the reactive group is activated or deactivated by electron beam lithography, x-ray lithography, or any other radiation. Suitable reactive groups for electron beam lithography include sulfonyl. Other methods may be used including, for example, exposure to a current source. Other reactive groups and methods of activation may be used in light of this disclosure.

As shown in FIG. 1, the linking molecules are preferably exposed to, for example, light through a suitable mask 8 using photolithographic techniques of the type known in the semiconductor industry and described in, for example, Sze, *VLSI Technology*, McGraw-Hill (1983), and Mead et al., *Introduction to VLSI Systems*, Addison-Wesley (1980), which are incorporated herein by reference for all purposes. The light may be directed at either the surface containing the protective groups or at the back of the substrate, so long as the substrate is transparent to the wavelength of light needed for removal of the protective groups. In the embodiment shown in FIG. 1, light is directed at the surface of the substrate containing the protective groups. FIG. 1 illustrates the use of such masking techniques as they are applied to a positive reactive group so as to activate linking molecules and expose functional groups in areas **10***a* and **10***b*.

The mask **8** is in one embodiment a transparent support material selectively coated with a layer of opaque material. Portions of the opaque material are removed, leaving opaque material in the precise pattern desired on the substrate surface. The mask is brought into close proximity with, imaged on, or brought directly into contact with the substrate surface as shown in FIG. 1. "Openings" in the mask correspond to locations on the substrate where it is desired to remove photoremovable protective groups from the substrate. Alignment may be performed using conventional alignment techniques in which alignment marks (not shown) are used to accurately overlay successive masks with previous patterning steps, or more sophisticated techniques may be used. For example, interferometric techniques such as the one described in Flanders et al., "A New Interferometric Alignment Technique," *App. Phys. Lett.* (1977) 31:426–428, which is incorporated herein by reference, may be used.

To enhance contrast of light applied to the substrate, it is desirable to provide contrast enhancement materials between the mask and the substrate according to some embodiments. This contrast enhancement layer may comprise a molecule which is decomposed by light such as quinone diazid or a material which is transiently bleached at the wavelength of interest. Transient bleaching of materials will allow greater penetration where light is applied, thereby enhancing contrast. Alternatively, contrast enhancement may be provided by way of a cladded fiber optic bundle.

The light may be from a conventional incandescent source, a laser, a laser diode, or the like. If non-collimated sources of light are used it may be desirable to provide a thick- or multi-layered mask to prevent spreading of the light onto the substrate. It may, further, be desirable in some embodiments to utilize groups which are sensitive to different wavelengths to control synthesis. For example, by using groups which are sensitive to different wavelengths, it is possible to select branch positions in the synthesis of a polymer or eliminate certain masking steps. Several reactive groups along with their corresponding wavelengths for deprotection are provided in Table 1.

TABLE 1

Group	Approximate Deprotection Wavelength
Nitroveratryloxy carbonyl (NVOC)	UV (300–400 nm)
Nitrobenzyloxy carbonyl (NBOC)	UV (300–350 nm)
Dimethyl dimethoxybenzyloxy carbonyl	UV (280–300 nm)
5-Bromo-7-nitroindolinyl	UV (420 nm)
o-Hydroxy-α-methyl cinnamoyl	UV (300–350 nm)
2-Oxymethylene anthraquinone	UV (350 nm)

While the invention is illustrated primarily herein by way of the use of a mask to illuminate selected regions the substrate, other techniques may also be used. For example, the substrate may be translated under a modulated laser or diode light source. Such techniques are discussed in, for example, U.S. Pat. No. 4,719,615 (Feyrer et al.), which is incorporated herein by reference. In alternative embodiments a laser galvanometric scanner is utilized. In other embodiments, the synthesis may take place on or in contact with a conventional liquid crystal (referred to herein as a "light valve") or fiber optic light sources. By appropriately modulating liquid crystals, light may be selectively controlled so as to permit light to contact selected regions of the substrate. Alternatively, synthesis may take place on the end of a series of optical fibers to which light is selectively applied. Other means of controlling the location of light exposure will be apparent to those of skill in the art.

The substrate may be irradiated either in contact or not in contact with a solution (not shown) and is, preferably, irradiated in contact with a solution. The solution contains reagents to prevent the by-products formed by irradiation from interfering with synthesis of the polymer according to some embodiments. Such by-products might include, for example, carbon dioxide, nitrosocarbonyl compounds, styrene derivatives, indole derivatives, and products of their photochemical reactions. Alternatively, the solution may contain reagents used to match the index of refraction of the substrate. Reagents added to the solution may further include, for example, acidic or basic buffers, thiols, substituted hydrazines and hydroxylamines, reducing agents (e.g., NADH) or reagents known to react with a given functional group (e.g., aryl nitroso+glyoxylic acid→aryl formhydroxamate+CO_2).

Either concurrently with or after the irradiation step, the linker molecules are washed or otherwise contacted with a first monomer, illustrated by "A" in regions **12***a* and **12***b* in FIG. 2. The first monomer reacts with the activated functional groups of the linkage molecules which have been exposed to light. The first monomer, which is preferably an amino acid, is also provided with a photoprotective group. The photoprotective group on the monomer may be the same as or different than the protective group used in the linkage molecules, and may be selected from any of the above-described protective groups. In one embodiment, the protective groups for the A monomer is selected from the group NBOC and NVOC.

As shown in FIG. 3, the process of irradiating is thereafter repeated, with a mask repositioned so as to remove linkage protective groups and expose functional groups in regions **14***a* and **14***b* which are illustrated as being regions which were protected in the previous masking step. As an alternative to repositioning of the first mask, in many embodiments a second mask will be utilized. In other alternative embodiments, some steps may provide for illuminating a common region in successive steps. As shown in FIG. 3, it may be desirable to provide separation between irradiated regions. For example, separation of about 1–5 μm may be appropriate to account for alignment tolerances.

As shown in FIG. 4, the substrate is then exposed to a second protected monomer "B," producing B regions **16***a* and **16***b*. Thereafter, the substrate is again masked so as to remove the protective groups and expose reactive groups on A region **12***a* and B region **16***b*. The substrate is again exposed to monomer B, resulting in the formation of the structure shown in FIG. 6. The dimers B-A and B-B have been produced on the substrate.

A subsequent series of masking and contacting steps similar to those described above with A (not shown) provides the structure shown in FIG. 7. The process provides all possible dimers of B and A, i.e., B-A, A-B, A-A, and B-B.

The substrate, the area of synthesis, and the area for synthesis of each individual polymer could be of any size or shape. For example, squares, ellipsoids, rectangles, triangles, circles, or portions thereof, along with irregular geometric shapes, may be utilized. Duplicate synthesis areas may also be applied to a single substrate for purposes of redundancy.

In one embodiment the regions **12** and **16** on the substrate will have a surface area of between about 1 cm^2 and 10^{-10}

cm^2. In some embodiments the regions **12** and **16** have areas of less than about 10^{-1} cm^2, 10^{-2} cm^2, 10^{-3} cm^2, 10^{-4} cm^2, 10^{-5} cm^2, 10^{-6} cm^2, 10^{-7} cm^2, 10^{-8} cm^2, or 10^{-10} cm^2. In a preferred embodiment, the regions **12** and **16** are between about 10×10 μm and 500×500 μm.

In some embodiments a single substrate supports more than about 10 different monomer sequences and perferably more than about 100 different monomer sequences, although in some embodiments more than about 10^3, 10^4, 10^5, 10^6, 10^7, or 10^8 different sequences are provided on a substrate. Of course, within a region of the substrate in which a monomer sequence is synthesized, it is preferred that the monomer sequence be substantially pure. In some embodiments, regions of the substrate contain polymer sequences which are at least about 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98%, or 99% pure.

According to some embodiments, several sequences are intentionally provided within a single region so as to provide an initial screening for biological activity, after which materials within regions exhibiting significant binding are further evaluated.

IV. Details of One Embodiment of a Reactor System

FIG. **8A** schematically illustrates a preferred embodiment of a reactor system **100** for synthesizing polymers on the prepared substrate in accordance with one aspect of the invention. The reactor system includes a body **102** with a cavity **104** on a surface thereof. In preferred embodiments the cavity **104** is between about 50 and 1000 μm deep with a depth of about 500 μm preferred.

The bottom of the cavity is preferably provided with an array of ridges **106** which extend both into the plane of the Figure and parallel to the plane of the Figure. The ridges are preferably about 50 to 200 μm deep and spaced at about 2 to 3mm. The purpose of the ridges is to generate turbulent flow for better mixing. The bottom surface of the cavity is preferably light absorbing so as to prevent reflection of impinging light.

A substrate **112** is mounted above the cavity **104**. The substrate is provided along its bottom surface **114** with a photoremovable protective group such as NVOC with or without an intervening linker molecule. The substrate is preferably transparent to a wide spectrum of light, but in some embodiments is transparent only at a wavelength at which the protective group may be removed (such as UV in the case of NVOC). The substrate in some embodiments is a conventional microscope glass slide or cover slip. The substrate is preferably as thin as possible, while still providing adequate physical support. Preferably, the substrate is less than about 1 mm thick, more preferably less than 0.5 mm thick, more preferably less than 0.1 mm thick, and most preferably less than 0.05 mm thick. In alternative preferred embodiments, the substrate is quartz or silicon.

The substrate and the body serve to seal the cavity except for an inlet port **108** and an outlet port **110**. The body and the substrate may be mated for sealing in some embodiments with one or more gaskets. According to a preferred embodiment, the body is provided with two concentric gaskets and the intervening space is held at vacuum to ensure mating of the substrate to the gaskets.

Fluid is pumped through the inlet port into the cavity by way of a pump **116** which may be, for example, a model no. B-120-S made by Eldex Laboratories. Selected fluids are circulated into the cavity by the pump, through the cavity, and out the outlet for recirculation or disposal. The reactor may be subjected to ultrasonic radiation and/or heated to aid in agitation in some embodiments.

Above the substrate **112**, a lens **120** is provided which may be, for example, a 2" 100 mm focal length fused silica lens. For the sake of a compact system, a reflective mirror **122** may be provided for directing light from a light source **124** onto the substrate. Light source **124** may be, for example, a Xe(Hg) light source manufactured by Oriel and having model no. 66024. A second lens **126** may be provided for the purpose of projecting a mask image onto the substrate in combination with lens **112**. This form of lithography is referred to herein as projection printing. As will be apparent from this disclosure, proximity printing and the like may also be used according to some embodiments.

Light from the light source is permitted to reach only selected locations on the substrate as a result of mask **128**. Mask **128** may be, for example, a glass slide having etched chrome thereon. The mask **128** in one embodiment is provided with a grid of transparent locations and opaque locations. Such masks may be manufactured by, for example, Photo Sciences, Inc. Light passes freely through the transparent regions of the mask, but is reflected from or absorbed by other regions. Therefore, only selected regions of the substrate are exposed to light.

As discussed above, light valves (LCD's) may be used as an alternative to conventional masks to selectively expose regions of the substrate. Alternatively, fiber optic faceplates such as those available from Schott Glass, Inc, may be used for the purpose of contrast enhancement of the mask or as the sole means of restricting the region to which light is applied. Such faceplates would be placed directly above or on the substrate in the reactor shown in FIG. **8A**. In still further embodiments, flys-eye lenses, tapered fiber optic faceplates, or the like, may be used for contrast enhancement.

In order to provide for illumination of regions smaller than a wavelength of light, more elaborate techniques may be utilized. For example, according to one preferred embodiment, light is directed at the substrate by way of molecular microcrystals on the tip of, for example, micropipettes. Such devices are disclosed in Lieberman et al., "A Light Source Smaller Than the Optical Wavelength," *Science* (1990) 247:59–61, which is incorporated herein by reference for all purposes.

In operation, the substrate is placed on the cavity and sealed thereto. All operations in the process of preparing the substrate are carried out in a room lit primarily or entirely by light of a wavelength outside of the light range at which the protective group is removed. For example, in the case of NVOC, the room should be lit with a conventional dark room light which provides little or no UV light. All operations are preferably conducted at about room temperature.

A first, deprotection fluid (without a monomer) is circulated through the cavity. The solution preferably is of 5 mM sulfuric acid in dioxane solution which serves to keep exposed amino groups protonated and decreases their reactivity with photolysis by-products. Absorptive materials such as N,N-diethylamino 2,4-dinitrobenzene, for example, may be included in the deprotection fluid which serves to absorb light and prevent reflection and unwanted photolysis.

The slide is, thereafter, positioned in a light raypath from the mask such that first locations on the substrate are illuminated and, therefore, deprotected. In preferred embodiments the substrate is illuminated for between about 1 and 15 minutes with a preferred illumination time of about 10 minutes at 10–20 mW/cm^2 with 365 nm light. The slides are neutralized (i.e., brought to a pH of about 7) after photolysis with, for example, a solution of di-isopropylethylamine (DIEA) in methylene chloride for about 5 minutes.

The first monomer is then placed at the first locations on the substrate. After irradiation, the slide is removed, treated in bulk, and then reinstalled in the flow cell. Alternatively, a fluid containing the first monomer, preferably also protected by a protective group, is circulated through the cavity by way of pump 116. If, for example, it is desired to attach the amino acid Y to the substrate at the first locations, the amino acid Y (bearing a protective group on its α-nitrogen), along with reagents used to render the monomer reactive, and/or a carrier, is circulated from a storage container 118, through the pump, through the cavity, and back to the inlet of the pump.

The monomer carrier solution is, in a preferred embodiment, formed by mixing of a first solution (referred to herein as solution "A") and a second solution (referred to herein as solution "B"). Table 2 provides an illustration of a mixture which may be used for solution A.

TABLE 2

Representative Monomer Carrier Solution "A"
100 mg NVOC amino protected amino acid
37 mg HOBT (1-Hydroxybenzotriazole)
250 μl DMF (Dimethylformamide)
86 μl DIEA (Diisopropylethylamine)

The composition of solution B is illustrated in Table 3. Solutions A and B are mixed and allowed to react at room temperature for about 8 minutes, then diluted with 2 ml of DMF, and 500 μl are applied to the surface of the slide or the solution is circulated through the reactor system and allowed to react for about 2 hours at room temperature. The slide is then washed with DMF, methylene chloride and ethanol.

TABLE 3

Representative Monomer Carrier Solution "B"
250 μl DXF
111 mg BOP (Benzotriazolyl-n-oxy-tris(dimethylamino) phosphoniumhexafluorophosphate)

As the solution containing the monomer to be attached is circulated through the cavity, the amino acid or other monomer will react at its carboxy terminus with amino groups on the regions of the substrate which have been deprotected. Of course, while the invention is illustrated by way of circulation of the monomer through the cavity, the invention could be practiced by way of removing the slide from the reactor and submersing it in an appropriate monomer solution.

After addition of the first monomer, the solution containing the first amino acid is then purged from the system. After circulation of a sufficient amount of the DMF/methylene chloride such that removal of the amino acid can be assured (e.g., about 50× times the volume of the cavity and carrier lines), the mask or substrate is repositioned, or a new mask is utilized such that second regions on the substrate will be exposed to light and the light 124 is engaged for a second exposure. This will deprotect second regions on the substrate and the process is repeated until the desired polymer sequences have been synthesized.

The entire derivatized substrate is then exposed to a receptor of interest, preferably labeled with, for example, a fluorescent marker, by circulation of a solution or suspension of the receptor through the cavity or by contacting the surface of the slide in bulk. The receptor will preferentially bind to certain regions of the substrate which contain complementary sequences.

Antibodies are typically suspended in what is commonly referred to as "supercocktail," which may be, for example, a solution of about 1% BSA (bovine serum albumin), 0.5% Tween in PBS (phosphate buffered saline) buffer. The antibodies are diluted into the supercocktail buffer to a final concentration of, for example, about 0.1 to 4 μg/ml.

FIG. 8B illustrates an alternative preferred embodiment of the reactor shown in FIG. 8A. According to this embodiment, the mask 128 is placed directly in contact with the substrate. Preferably, the etched portion of the mask is placed face down so as to reduce the effects of light dispersion. According to this embodiment, the imaging lenses 120 and 126 are not necessary because the mask is brought into close proximity with the substrate.

For purposes of increasing the signal-to-noise ratio of the technique, some embodiments of the invention provide for exposure of the substrate to a first labeled or unlabeled receptor followed by exposure of a labeled, second receptor (e.g., an antibody) which binds at multiple sites on the first receptor. If, for example, the first receptor is an antibody derived from a first species of an animal, the second receptor is an antibody derived from a second species directed to epitopes associated with the first species. In the case of a mouse antibody, for example, fluorescently labeled goat antibody or antiserum which is antimouse may be used to bind at multiple sites on the mouse antibody, providing several times the fluorescence compared to the attachment of a single mouse antibody at each binding site. This process may be repeated again with additional antibodies (e.g., goat-mouse-goat, etc.) for further signal amplification.

In preferred embodiments an ordered sequence of masks is utilized. In some embodiments it is possible to use as few as a single mask to synthesize all of the possible polymers of a given monomer set.

If, for example, it is desired to synthesize all 16 dinucleotides from four bases, a 1 cm square synthesis region is divided conceptually into 16 boxes, each 0.25 cm wide. Denote the four monomer units by A, B, C, and D. The first reactions are carried out in four vertical columns, each 0.25 cm wide. The first mask exposes the left-most column of boxes, where A is coupled. The second mask exposes the next column, where B is coupled; followed by a third mask, for the C column; and a final mask that exposes the right-most column, for D. The first, second, third, and fourth masks may be a single mask translated to different locations.

The process is repeated in the horizontal direction for the second unit of the dimer. This time, the masks allow exposure of horizontal rows, again 0.25 cm wide. A, B, C, and D are sequentially coupled using masks that expose horizontal fourths of the reaction area. The resulting substrate contains all 16 dinucleotides of four bases.

The eight masks used to synthesize the dinucleotide are related to one another by translation or rotation. In fact, one mask can be used in all eight steps if it is suitably rotated and translated. For example, in the example above, a mask with a single transparent region could be sequentially used to expose each of the vertical columns, translated 90°, and then sequentially used to allow exposure of the horizontal rows.

Tables 4 and 5 provide a simple computer program in Quick Basic for planning a masking program and a sample output, respectively, for the synthesis of a polymer chain of three monomers ("residues") having three different monomers in the first level, four different monomers in the second level, and five different monomers in the third level in a striped pattern. The output of the program is the number of cells, the number of "stripes" (light regions) on each mask, and the amount of translation required for each exposure of the mask.

TABLE 4

Mask Strategy Program

```
DEFINT A-Z
DIM b(20), w(20), 1(500)
F$ - "LPT1:"
OPEN £S FOR OUTPUT AS #1
jmax = 3             'Number of residues
b(1) = 3: b(2) = 4: b(3) = 5          'Number of building blocks for res 1, 2, 3
g = 1: 1max(1) = 1
FOR j = 1 TO jmax: g= g * b(j): NEXT j
w(0) = 0: w(1) = g/b(1)
PRINT #1, "MASK2.BAS", DATE$, TIME$: PRINT #1,
PRINT #1, USING "Number of residues=##"; jmax
FOR j = 1 TO jmax
PRINT #1, USING "      Residue ##           ## building blocks"; j; b(j)
NEXT j
PRINT #1, "
PRINT #1, USING "Number of cells=####"; g: PRINT #1,
FOR j = 2 TO jmax
1max(j) = 1max(j - 1) * b(j - 1)
w(j) = w(j - 1) / b(j)
NEXT j
FOR j = 1 TO jmax
PRINT #1, USING "Mask for residue ##"; j: PRINT #1,
PRINT #1, USING "      Number of stripes=###"; 1max(j)
PRINT #1, USING "      Width of each stripe=###"; w(j)
FOR 1 = 1 TO 1max(j)
a = 1 + (1 - 1) * w(j - 1)
ae = a + w(j) - 1
PRINT #1, USING "      Stripe ## begins at location ### and ends at ###"; 1; a; ae
NEXT 1
PRINT #1,
PRINT #1, USING "      For each of ## building blocks, translate mask by ##
cell(s)"; b(j); w(j),
PRINT #1, : PRINT #1, : PRINT #1,
NEXT j
```

TABLE 5

Masking Strategy Output

```
Number of residues= 3

  Residue 1            3 building blocks
  Residue 2            4 building blocks
  Residue 3            5 building blocks
Number of cells= 60
Mask for residue 1

  Number of stripes= 1
  Width of each stripe= 20
  Stripe 1 begins at location 1 and ends at 20
  For each of 3 building blocks, translate mask by 20 cell(s)
Mask for residue 2

  Number of stripes= 3
  Width of each stripe= 5
  Stripe 1 begins at location 1 and ends at 5
  Stripe 2 begins at location 21 and ends at 25
  Stripe 3 begins at location 41 and ends at 45
  For each of 4 building blocks, translate mask by 5 cell(s)
Mask for residue 3

Number of stripes= 12
Width of each stripe= 1
Stripe 1 begins at location 1 and ends at 1
Stripe 2 begins at location 6 and ends at 6
Stripe 3 begins at location 11 and ends at 11
Stripe 4 begins at location 16 and ends at 16
```

TABLE 5-continued

Masking Strategy Output
Stripe 5 begins at location 21 and ends at 21
Stripe 6 begins at location 26 and ends at 26
Stripe 7 begins at location 31 and ends at 31
Stripe 8 begins at location 36 and ends at 36
Stripe 9 begins at location 41 and ends at 41
Stripe 10 begins at location 46 and ends at 46
Stripe 11 begins at location 51 and ends at 51
Stripe 12 begins at location 56 and ends at 56
For each of 5 building blocks, translate mask by 1 cell(s)

V. Details of One Embodiment of A Fluorescent Detection Device

FIG. 9 illustrates a fluorescent detection device for detecting fluorescently labeled receptors on a substrate. A substrate **112** is placed on an x/y translation table **202**. In a preferred embodiment the x/y translation table is a model no. PM500-A1 manufactured by Newport Corporation. The x/y translation table is connected to and controlled by an appropriately programmed digital computer **204** which may be, for example, an appropriately programmed IBM PC/AT or AT compatible computer. Of course, other computer systems, special purpose hardware, or the like could readily be substituted for the AT computer used herein for illustration. Computer software for the translation and data collection functions described herein can be provided based on commercially available software including, for example, "Lab Windows" licensed by National Instruments, which is incorporated herein by reference for all purposes.

The substrate and x/y translation table are placed under a microscope **206** which includes one or more objectives **208**. Light (about 488 nm) from a laser **210**, which in some embodiments is a model no. 2020-05 argon ion laser manufactured by Spectraphysics, is directed at the substrate by a dichroic mirror **207** which passes greater than about 520 nm light but reflects 488 nm light. Dichroic mirror **207** may be, for example, a model no. FT510 manufactured by Carl Zeiss. Light reflected from the mirror then enters the microscope **206** which may be, for example, a model no. Axioscop 20 manufactured by Carl Zeiss. Fluorescein-marked materials on the substrate will fluoresce >488 nm light, and the fluoresced light will be collected by the microscope and passed through the mirror. The fluorescent light from the substrate is then directed through a wavelength filter **209** and, thereafter through an aperture plate **211**. Wavelength filter **209** may be, for example, a model no. OG530 manufactured by Melles Griot and aperture plate **211** may be, for example, a model no. 477352/477380 manufactured by Carl Zeiss.

The fluoresced light then enters a photomultiplier tube **212** which in some embodiments is a model no. R943-02 manufactured by Hamamatsu, the signal is amplified in preamplifier **214** and photons are counted by photon counter **216**. The number of photons is recorded as a function of the location in the computer **204**. Pre-Amp **214** may be, for example, a model no. SR440 manufactured by Stanford Research Systems and photon counter **216** may be a model no. SR400 manufactured by Stanford Research Systems. The substrate is then moved to a subsequent location and the process is repeated. In preferred embodiments the data are acquired every 1 to 100 μm with a data collection diameter of about 0.8 to 10 μm preferred. In embodiments with sufficiently high fluorescence, a CCD detector with broadfield illumination is utilized.

By counting the number of photons generated in a given area in response to the laser, it is possible to determine where fluorescent marked molecules are located on the substrate. Consequently, for a slide which has a matrix of polypeptides, for example, synthesized on the surface thereof, it is possible to determine which of the polypeptides is complementary to a fluorescently marked receptor.

According to preferred embodiments, the intensity and duration of the light applied to the substrate is controlled by varying the laser power and scan stage rate for improved signal-to-noise ratio by maximizing fluorescence emission and minimizing background noise.

While the detection apparatus has been illustrated primarily herein with regard to the detection of marked receptors, the invention will find application in other areas. For example, the detection apparatus disclosed herein could be used in the fields of catalysis, DNA or protein gel scanning, and the like.

VI. Determination of Relative Binding Strength of Receptors

The signal-to-noise ratio of the present invention is sufficiently high that not only can the presence or absence of a receptor on a ligand be detected, but also the relative binding affinity of receptors to a variety of sequences can be determined.

In practice it is found that a receptor will bind to several peptide sequences in an array, but will bind much more strongly to some sequences than others. Strong binding affinity will be evidenced herein by a strong fluorescent or radiographic signal since many receptor molecules will bind in a region of a strongly bound ligand. Conversely, a weak binding affinity will be evidenced by a weak fluorescent or radiographic signal due to the relatively small number of receptor molecules which bind in a particular region of a substrate having a ligand with a weak binding affinity for the receptor. consequently, it becomes possible to determine relative binding avidity (or affinity in the case of univalent interactions) of a ligand herein by way of the intensity of a fluorescent or radiographic signal in a region containing that ligand.

Semiquantitative data on affinities might also be obtained by varying washing conditions and concentrations of the receptor. This would be done by comparison to known ligand receptor pairs, for example.

VII. Examples

The following examples are provided to illustrate the efficacy of the inventions herein. All operations were conducted at about ambient temperatures and pressures unless indicated to the contrary.

A. Slide Preparation

Before attachment of reactive groups it is preferred to clean the substrate which is, in a preferred embodiment a glass substrate such as a microscope slide or cover slip.

According to one embodiment the slide is soaked in an alkaline bath consisting of, for example, 1 liter of 95% ethanol with 120 ml of water and 120 grams of sodium hydroxide for 12 hours. The slides are then washed under running water and allowed to air dry, and rinsed once with a solution of 95% ethanol.

The slides are then aminated with, for example, aminopropyltriethoxysilane for the purpose of attaching amino groups to the glass surface on linker molecules, although any omega functionalized silane could also be used for this purpose. In one embodiment 0.1% aminopropyltriethoxysilane is utilized, although solutions with concentrations from 10^{-7}% to 10% may be used, with about 10^{-3}% to 2% preferred. A 0.1% mixture is prepared by adding to 100 ml of a 95% ethanol/5% water mixture, 100 microliters (μl) of aminopropyltriethoxysilane. The mixture is agitated at about ambient temperature on a rotary shaker for about 5 minutes. 500 μl of this mixture is then applied to the surface of one side of each cleaned slide. After 4 minutes, the slides are decanted of this solution and rinsed three times by dipping in, for example, 100% ethanol.

After the plates dry, they are placed in a 110–120° C. vacuum oven for about 20 minutes, and then allowed to cure at room temperature for about 12 hours in an argon environment. The slides are then dipped into DMF (dimethylformamide) solution, followed by a thorough washing with methylene chloride.

The aminated surface of the slide is then exposed to about 500 μl of, for example, a 30 millimolar (mM) solution of NVOC-GABA (gamma amino butyric acid) NHS (N-hydroxysuccinimide) in DMF for attachment of a NVOC-GABA to each of the amino groups.

The surface is washed with, for example, DMF, methylene chloride, and ethanol.

Any unreacted aminopropyl silane on the surface—that is, those amino groups which have not had the NVOC-GABA attached—are now capped with acetyl groups (to prevent further reaction) by exposure to a 1:3 mixture of acetic anhydride in pyridine for 1 hour. Other materials which may perform this residual capping function include trifluoroacetic anhydride, formicacetic anhydride, or other reactive acylating agents. Finally, the slides are washed again with DMF, methylene chloride, and ethanol.

B. Synthesis of Eight Trimers of "A" and "B"

FIG. 10 illustrates a possible synthesis of the eight trimers of the two-monomer set: gly, phe (represented by "A" and "B," respectively). A glass slide bearing silane groups terminating in 6-nitroveratryloxycarboxamide (NVOC-NH) residues is prepared as a substrate. Active esters (pentafluorophenyl, OBt, etc.) of gly and phe protected at the amino group with NVOC are prepared as reagents. While not pertinent to this example, if side chain protecting groups are required for the monomer set, these must not be photoreactive at the wavelength of light used to protect the primary chain.

For a monomer set of size n, nxl cycles are required to synthesize all possible sequences of length l. A cycle consists of:

1. Irradiation through an appropriate mask to expose the amino groups at the sites where the next residue is to be added, with appropriate washes to remove the by-products of the deprotection.
2. Addition of a single activated and protected (with the same photochemically-removable group) monomer, which will react only at the sites addressed in step 1, with appropriate washes to remove the excess reagent from the surface.

The above cycle is repeated for each member of the monomer set until each location on the surface has been extended by one residue in one embodiment. In other embodiments, several residues are sequentially added at one location before moving on to the next location. Cycle times will generally be limited by the coupling reaction rate, now as short as 20 min in automated peptide synthesizers. This step is optionally followed by addition of a protecting group to stabilize the array for later testing. For some types of polymers (e.g., peptides), a final deprotection of the entire surface (removal of photoprotective side chain groups) may be required.

More particularly, as shown in FIG. **10**A, the glass **20** is provided with regions **22**, **24**, **26**, **28**, **30**, **32**, **34**, and **36**. Regions **30**, **32**, **34**, and **36** are masked, as shown in FIG. **10**B and the glass is irradiated and exposed to a reagent containing "A" (e.g., gly), with the resulting structure shown in FIG. **10**C. Thereafter, regions **22**, **24**, **26**, and **28** are masked, the glass is irradiated (as shown in FIG. **10**D) and exposed to a reagent containing "B" (e.g., phe), with the resulting structure shown in FIG. **10**E. The process proceeds, consecutively masking and exposing the sections as shown until the structure shown in FIG. **10**M is obtained. The glass is irradiated and the terminal groups are, optionally, capped by acetylation. As shown, all possible trimers of gly/phe are obtained.

In this example, no side chain protective group removal is necessary. If it is desired, side chain deprotection may be accomplished by treatment with ethanedithiol and trifluoroacetic acid.

In general, the number of steps needed to obtain a particular polymer chain is defined by:

$$n \times l \qquad (1)$$

where:

n=the number of monomers in the basis set of monomers, and

l=the number of monomer units in a polymer chain.

Conversely, the synthesized number of sequences of length l will be:

$$n^l. \qquad (2)$$

Of course, greater diversity is obtained by using masking strategies which will also include the synthesis of polymers having a length of less than l. If, in the extreme case, all polymers having a length less than or equal to l are synthesized, the number of polymers synthesized will be:

$$n^l+n^{l-1}+\ldots+n^1. \qquad (3)$$

The maximum number of lithographic steps needed will generally be n for each "layer" of monomers, i.e., the total number of masks (and, therefore, the number of lithographic steps) needed will be nxl. The size of the transparent mask regions will vary in accordance with the area of the substrate available for synthesis and the number of sequences to be formed. In general, the size of the synthesis areas will be:

$$\text{size of synthesis areas}=(A)/(S)$$

where:

A is the total area available for synthesis; and

S is the number of sequences desired in the area.

It will be appreciated by those of skill in the art that the above method could readily be used to simultaneously produce thousands or millions of oligomers on a substrate using the photolithographic techniques disclosed herein. Consequently, the method results in the ability to practically test large numbers of, for example, di, tri, tetra, penta, hexa, hepta, octapeptides, dodecapeptides, or larger polypeptides (or correspondingly, polynucleotides).

The above example has illustrated the method by way of a manual example. It will of course be appreciated that automated or semi-automated methods could be used. The substrate would be mounted in a flow cell for automated addition and removal of reagents, to minimize the volume of reagents needed, and to more carefully control reaction conditions. Successive masks could be applied manually or automatically.

C. Synthesis of a Dimer of an Aminopropyl Group and a Fluorescent Group

In synthesizing the dimer of an aminopropyl group and a fluorescent group, a functionalized durapore membrane was used as a substrate. The durapore membrane was a polyvinylidine difluoride with aminopropyl groups. The aminopropyl groups were protected with the DDZ group by reaction of the carbonyl chloride with the amino groups, a reaction readily known to those of skill in the art. The surface bearing these groups was placed in a solution of THF and contacted with a mask bearing a checkerboard pattern of 1 mm opaque and transparent regions. The mask was exposed to ultraviolet light having a wavelength down to at least about 280 nm for about 5 minutes at ambient temperature, although a wide range of exposure times and temperatures may be appropriate in various embodiments of the invention. For example, in one embodiment, an exposure time of between about 1 and 5000 seconds may be used at process temperatures of between −70 and +50° C.

In one preferred embodiment, exposure times of between about 1 and 500 seconds at about ambient pressure are used. In some preferred embodiments, pressure above ambient is used to prevent evaporation.

The surface of the membrane was then washed for about 1 hour with a fluorescent label which included an active ester bound to a chelate of a lanthanide. Wash times will vary over a wide range of values from about a few minutes to a few hours. These materials fluoresce in the red and the green visible region. After the reaction with the active ester in the fluorophore was complete, the locations in which the fluorophore was bound could be visualized by exposing them to ultraviolet light and observing the red and the green fluorescence. It was observed that the derivatized regions of the substrate closely corresponded to the original pattern of the mask.

D. Demonstration of Signal Capability

Signal detection capability was demonstrated using a low-level standard fluorescent bead kit manufactured by Flow Cytometry Standarda and having model no. 824. This kit includes 5.8 μm diameter beads, each impregnated with a known number of fluorescein molecules.

One of the beads was placed in the illumination field on the scan stage as shown in FIG. 9 in a field of a laser spot which was initially shuttered. After being positioned in the illumination field, the photon detection equipment was turned on. The laser beam was unblocked and it interacted with the particle bead, which then fluoresced. Fluorescence curves of beads impregnated with 7,000; 13,000; and 29,000 fluorescein molecules, are shown in FIGS. 11A, 11B, and 11C respectively. On each curve, traces for beads without fluorescein molecules are also shown. These experiments were performed with 488 nm excitation, with 100 μW of laser power. The light was focused through a 40 power 0.75 NA objective.

The fluorescence intensity in all cases started off at a high value and then decreased exponentially. The fall-off in intensity is due to photobleaching of the fluorescein molecules. The traces of beads without fluorescein molecules are used for background subtraction. The difference in the initial exponential decay between labeled and nonlabeled beads is integrated to give the total number of photon counts, and this number is related to the number of molecules per bead. Therefore, it is possible to deduce the number of photons per fluorescein molecule that can be detected. For the curves illustrated in FIG. 11, this calculation indicates the radiation of about 40 to 50 photons per fluorescein molecule are detected.

E. Determination of the Number of Molecules Per Unit Area

Aminopropylated glass microscope slides prepared according to the methods discussed above were utilized in order to establish the density of labeling of the slides. The free amino termini of the slides were reacted with FITC (fluorescein isothiocyanate) which forms a covalent linkage with the amino group. The slide is then scanned to count the number of fluorescent photons generated in a region which, using the estimated 40–50 photons per fluorescent molecule, enables the calculation of the number of molecules which are on the surface per unit area.

A slide with aminopropyl silane on its surface was immersed in a 1 mM solution of FITC in DMF for 1 hour at about ambient temperature. After reaction, the slide was washed twice with DMF and then washed with ethanol, water, and then ethanol again. It was then dried and stored in the dark until it was ready to be examined.

Through the use of curves similar to those shown in FIG. 11, and by integrating the fluorescent counts under the exponentially decaying signal, the number of free amino groups on the surface after derivitization was determined. It was determined that slides with labeling densities of 1 fluoroscein per $10^3 \times 10^3$ to ~2×2 nm could be reproducibly made as the concentration of aminopropyltriethoxysilane varied from 10^{-5}% to 10^{-1}%.

F. Removal of NVOC and Attachment of a Fluorescent Marker

NVOC-GABA groups were attached as described above. The entire surface of one slide was exposed to light so as to expose a free amino group at the end of the gamma amino butyric acid. This slide, and a duplicate which was not exposed, were then exposed to fluorescein isothiocyanate (FITC).

FIG. 12A illustrates the slide which was not exposed to light, but which was exposed to FITC. The units of the x axis are time and the units of the y axis are counts. The trace contains a certain amount of background fluorescence. The duplicate slide was exposed to 350 nm broadband illumination for about 1 minute (12 mW/cm^2, ~350 nm illumination), washed and reacted with FITC. The fluorescence curves for this slide are shown in FIG. 12B. A large increase in the level of fluorescence is observed, which indicates photolysis has exposed a number of amino groups on the surface of the slides for attachment of a fluorescent marker.

G. Use of a Mask in Removal of NVOC

The next experiment was performed with a 0.1% aminopropylated slide. Light from a Hg—Xe arc lamp was imaged onto the substrate through a laser-ablated chrome-on-glass mask in direct contact with the substrate.

This slide was illuminated for approximately 5 minutes, with 12 mW of 350 nm broadband light and then reacted with the 1 mM FITC solution. It was put on the laser detection scanning stage and a graph was plotted as a two-dimensional representation of position versus fluorescence intensity. The fluorescence intensity (in counts) as a function of location is given on the scale to the right of FIG. 13A for a mask having 100×100 μm squares.

The experiment was repeated a number of times through various masks. The fluorescence pattern for a 50 μm mask is illustrated in FIG. 13B, for a 20 μm mask in FIG. 13C, and for a 10 μm mask in FIG. 13D. The mask pattern is distinct down to at least about 10 μm squares using this lithographic technique.

H. Attachment of YGGFL and Subsequent Exposure to Herz Antibody and Goat Antimouse

In order to establish that receptors to a particular polypeptide sequence would bind to a surface-bound peptide and be detected, Leu enkephalin was coupled to the surface and recognized by an antibody. A slide was derivatized with 0.1% amino propyl-triethoxysilane and protected with NVOC. A 500 μm checkerboard mask was used to expose the slide in a flow cell using backside contact printing. The Leu enkephalin sequence (H_2N-tyrosine,glycine,glycine, phenylalanine,leucine-CO_2H, otherwise referred to herein as YGGFL) was attached via its carboxy end to the exposed amino groups on the surface of the slide. The peptide was added in DMF solution with the BOP/HOBT/DIEA coupling reagents and recirculated through the flow cell for 2 hours at room temperature.

A first antibody, known as the Herz antibody, was applied to the surface of the slide for 45 minutes at 2 μg/ml in a supercocktail (containing 1% BSA and 1% ovalbumin also in this case). A second antibody, goat anti-mouse fluorescein conjugate, was then added at 2 μg/ml in the supercocktail buffer, and allowed to incubate for 2 hours.

The results of this experiment are provided in FIG. 14. Again, this figure illustrates fluorescence intensity as a function of position. The fluorescence scale is shown on the right. This image was taken at 10 μm steps. This figure indicates that not only can deprotection be carried out in a well defined pattern, but also that (1) the method provides for successful coupling of peptides to the surface of the substrate, (2) the surface of a bound peptide is available for binding with an antibody, and (3) that the detection apparatus capabilities are sufficient to detect binding of a receptor.

I. Monomer-by-Monomer Formation of YGGFL and Subsequent Exposure to Labeled Antibody

Monomer-by-monomer synthesis of YGGFL and GGFL in alternate squares was performed on a slide in a checkerboard pattern and the resulting slide was exposed to the Herz antibody. This experiment and the results thereof are illustrated in FIGS. 15A, 15B, 15C, and 15D.

In FIG. 15A, a slide is shown which is derivatized with the aminopropyl group, protected in this case with t-BOC (t-butoxycarbonyl). The slide was treated with TFA to remove the t-BOC protecting group. E-aminocaproic acid, which was t-BOC protected at its amino group, was then coupled onto the aminopropyl groups. The aminocaproic acid serves as a spacer between the aminopropyl group and the peptide to be synthesized. The amino end of the spacer was deprotected and coupled to NVOC-leucine. The entire slide was then illuminated with 12 mW of 325 nm broadband illumination. The slide was then coupled with NVOC-phenylalanine and washed. The entire slide was again illuminated, then coupled to NVOC-glycine and washed. The slide was again illuminated and coupled to NVOC-glycine to form the sequence shown in the last portion of FIG. 15A.

As shown in FIG. 15B, alternating regions of the slide were then illuminated using a projection print using a 500×500 μm checkerboard mask; thus, the amino group of glycine was exposed only in the lighted areas. When the next coupling chemistry step was carried out, NVOC-tyrosine was added, and it coupled only at those is spots which had received illumination. The entire slide was then illuminated to remove all the NVOC groups, leaving a checkerboard of YGGFL in the lighted areas and in the other areas, GGFL. The Herz antibody (which recognizes the YGGFL, but not GGFL) was then added, followed by goat anti-mouse fluorescein conjugate.

The resulting fluorescence scan is shown in FIG. 15C, and the scale for the fluorescence intensity is again given on the right. Dark areas contain the tetrapeptide GGFL, which is not recognized by the Herz antibody (and thus there is no binding of the goat anti-mouse antibody with fluorescein conjugate), and in the red areas YGGFL is present. The YGGFL pentapeptide is recognized by the Herz antibody and, therefore, there is antibody in the lighted regions for the fluorescein-conjugated goat anti-mouse to recognize.

Similar patterns are shown for a 50 μm mask used in direct contact ("proximity print") with the substrate in FIG. 15D. Note that the pattern is more distinct and the corners of the checkerboard pattern are touching when the mask is placed in direct contact with the substrate (which reflects the increase in resolution using this technique).

J. Monomer-by-Monomer Synthesis of YGGFL and PGGFL

A synthesis using a 50 μm checkerboard mask similar to that shown in FIG. 15 was conducted. However, P was added to the GGFL sites on the substrate through an additional coupling step. P was added by exposing protected GGFL to light through a mask, and subsequence exposure to P in the manner set forth above. Therefore, half of the regions on the substrate contained YGGFL and the remaining half contained PGGFL.

The fluorescence plot for this experiment is provided in FIG. 16. As shown, the regions are again readily discernable. This experiment demonstrates that antibodies are able to recognize a specific sequence and that the recognition is not length-dependent.

K. Monomer-by-Monomer Synthesis of YGGFL and YPGGFL

In order to further demonstrate the operability of the invention, a 50 μm checkerboard pattern of alternating YGGFL and YPGGFL was synthesized on a substrate using techniques like those set forth above. The resulting fluorescence plot is provided in FIG. 17. Again, it is seen that the antibody is clearly able to recognize the YGGFL sequence and does not bind significantly at the YPGGFL regions.

L. Synthesis of an Array of Sixteen Different Amino Acid Sequences and Estimation of Relative Binding Affinity to Herz Antibody

Using techniques similar to those set forth above, an array of 16 different amino acid sequences (replicated four times) was synthesized on each of two glass substrates. The sequences were synthesized by attaching the sequence NVOC-GFL across the entire surface of the slides. Using a series of masks, two layers of amino acids were then selectively applied to the substrate. Each region had dimensions of 0.25 cm×0.0625 cm. The first slide contained amino acid sequences containing only L amino acids while the second slide contained selected D amino acids. FIGS. 18A and 18B illustrate a map of the various regions on the first and second slides, respectively. The patterns shown in FIGS. 18A and 18B were duplicated four times on each slide. The slides were then exposed to the Herz antibody and fluorescein-labeled goat anti-mouse.

FIG. 19 is a fluorescence plot of the first slide, which contained only L amino acids. Red indicates strong binding (149,000 counts or more) while black indicates little or no binding of the Herz antibody (20,000 counts or less). The bottom right-hand portion of the slide appears "cut off" because the slide was broken during processing. The sequence YGGFL is clearly most strongly recognized. The sequences YAGFL and YSGFL also exhibit strong recognition of the antibody. By contrast, most of the remaining sequences show little or no binding. The four duplicate portions of the slide are extremely consistent in the amount of binding shown therein.

FIG. 20 is a fluorescence plot of the second slide. Again, strongest binding is exhibited by the YGGFL sequence. Significant binding is also detected to YaGFL, YsGFL, and YpGFL. The remaining sequences show less binding with the antibody. Note the low binding efficiency of the sequence yGGFL.

Table 6 lists the various sequences tested in order of relative fluorescence, which provides information regarding relative binding affinity.

TABLE 6

Apparent Binding to Herz Ab	
L-a.a. Set	D-a.a. Set
YGGFL	YGGFL
YAGFL	YaGFL
YSGFL	YsGFL
LGGFL	YpGFL
FGGFL	fGGFL
YPGFL	yGGFL
LAGFL	faGFL
FAGFL	WGGFL
WGGFL	yaGFL
	fpGFL
	WaGFL

VIII. Illustrative Alternative Embodiment

According to an alternative embodiment of the invention, the methods provide for attaching to the surface a caged binding member which in its caged form has a relatively low affinity for other potentially binding species, such as receptors and specific binding substances. Such techniques are more fully described in copending application Ser. No. 404,920, filed Sep. 8, 1989, and incorporated herein by reference for all purposes.

According to this alternative embodiment, the invention provides methods for forming predefined regions on a surface of a solid support, wherein the predefined regions are capable of immobilizing receptors. The methods make use of caged binding members attached to the surface to enable selective activation of the predefined regions. The caged binding members are liberated to act as binding members ultimately capable of binding receptors upon selective activation of the predefined regions. The activated binding members are then used to immobilize specific molecules such as receptors on the predefined region of the surface. The above procedure is repeated at the same or different sites on the surface so as to provide a surface prepared with a plurality of regions on the surface containing, for example, the same or different receptors. When receptors immobilized in this way have a differential affinity for one or more ligands, screenings and assays for the ligands can be conducted in the regions of the surface containing the receptors.

The alternative embodiment may make use of novel caged binding members attached to the substrate. Caged (unactivated) members have a relatively low affinity for receptors of substances that specifically bind to uncaged binding members when compared with the corresponding affinities of activated binding members. Thus, the binding members are protected from reaction until a suitable source of energy is applied to the regions of the surface desired to be activated. Upon application of a suitable energy source, the caging groups labilize, thereby presenting the activated binding member. A typical energy source will be light.

Once the binding members on the surface are activated they may be attached to a receptor. The receptor chosen may be a monoclonal antibody, a nucleic acid sequence, a drug receptor, etc. The receptor will usually, though not always, be prepared so as to permit attaching it, directly or indirectly, to a binding member. For example, a specific binding substance having a strong binding affinity for the binding member and a strong affinity for the receptor or a conjugate of the receptor may be used to act as a bridge between binding members and receptors if desired. The method uses a receptor prepared such that the receptor retains its activity toward a particular ligand.

Preferably, the caged binding member attached to the solid substrate will be a photoactivatable biotin complex, i.e., a biotin molecule that has been chemically modified with photoactivatable protecting groups so that it has a significantly reduced binding affinity for avidin or avidin analogs than does natural biotin. In a preferred embodiment, the protecting groups localized in a predefined region of the surface will be removed upon application of a suitable source of radiation to give binding members, that are biotin or a functionally analogous compound having substantially the same binding affinity for avidin or avidin analogs as does biotin.

In another preferred embodiment, avidin or an avidin analog is incubated with activated binding members on the surface until the avidin binds strongly to the binding members. The avidin so immobilized on predefined regions of the surface can then be incubated with a desired receptor or conjugate of a desired receptor. The receptor will preferably be biotinylated, e.g., a biotinylated antibody, when avidin is immobilized on the predefined regions of the surface. Alternatively, a preferred embodiment will present an avidin/biotinylated receptor complex, which has been previously prepared, to activated binding members on the surface.

IX. Conclusion

The present inventions provide greatly improved methods and apparatus for synthesis of polymers on substrates. It is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments will be apparent to those of skill in the art upon reviewing the above description. By way of example, the invention has been described primarily with reference to the use of photoremovable protective groups, but it will be readily recognized by those of skill in the art that sources of radiation other than light could also be used. For example, in some embodiments it may be desirable to use protective groups which are sensitive to electron beam irradiation, x-ray irradiation, in combination with electron beam lithograph, or x-ray lithography techniques. Alternatively, the group could be removed by exposure to an electric current. The scope of the invention should, therefore, be determined not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.

What is claimed is:

1. An array of oligonucleotides, the array comprising:
   - a planar solid support having at least a first surface; and
   - a plurality of different oligonucleotides attached to the first surface of the solid support at a density exceeding 400 different oligonucleotides/cm^2, wherein each of the different oligonucleotides is attached to the surface of the solid support in a different known location, and has a different determinable sequence.

2. The array of claim 1, wherein each different oligonucleotides is from about 4 to about 20 nucleotides in length.

3. The array of claim 1, wherein each different oligonucleotide is at least 12 nucleotides in length.

4. The array of claim 1, wherein each different oligonucleotide is 2–100 nucleotides in length.

5. The array of claim 1, wherein the array comprises at least 1,000 different oligonucleotides attached to the first surface of the solid support.

6. The array of claim 1, wherein the array comprises at least 10,000 different oligonucleotides attached to the first surface of the solid support.

7. The array of claim 1, wherein each of the different known locations is physically separated from each other of the known locations.

8. The array of claim 1, wherein said planar solid support is glass.

9. The array of claim 1, wherein said oligonucleotides are attached to the first surface of the solid support through a linker group.

10. The array of claim 1, wherein the oligonucleotide in the different known locations are at least 20% pure.

11. The array of claim 1, wherein the oligonucleotides in the different known locations are at least 50% pure.

12. The array of claim 1, wherein the oligonucleotide in the different known locations are at least 80% pure.

13. The array of claim 1, wherein the oligonucleotide in the different known locations are at least 90% pure.

14. The array of claim 1, wherein said array is produced by a binary synthesis process, said process comprising the steps of:

providing a planar solid support, said solid support having a plurality of compounds immobilized on a surface thereof, said compounds having protecting groups coupled thereto;

deprotecting a first portion of said plurality of compounds on said surface and not a second portion of said plurality of compounds;

reacting said first portion of said plurality of compounds with a first component of said oligonucleotide;

deprotecting at least a third portion of said plurality of compounds on said surface, said third portion comprising a fraction of said first portion of said plurality of compounds;

reacting said at least third portion of said plurality of compounds with a second component of said oligonucleotide; and

optionally repeating said binary synthesis steps to produce said oligonucleotide array.

15. An array of nucleic acids, the array comprising:

a planar support having at least a first surface; and

a plurality of different nucleic acids attached to the first surface of the solid support at a density exceeding 400 different nucleic acids/cm^2, wherein each of the different nucleic acids is attached to the surface of the solid support in a different known location, has a different determinable sequence, wherein the different nucleic acids in the different known locations are at least 10% pure.

16. The array of claim 15, wherein each different nucleic acid is at least 20 nucleotides in length.

17. The array of claim 15, wherein the array comprises at least 1,000 different nucleic acids attached to the first surface of the solid support.

18. The array of claim 15, wherein the array comprises at least 10,000 different nucleic acids attached to the first surface of the solid support.

19. The array of claim 15, wherein each of the different known locations is physically separated from each of the other known locations.

20. The array of claim 15, wherein said planar solid support is glass.

21. The array of claim 15, wherein said nucleic acids are attached to the first surface of the solid support through a linker group.

22. The array of claim 15, wherein the nucleic acids in the different known locations comprise nucleic acids that are at least 20% pure.

23. The array of claim 15, wherein the nucleic acid in the different known locations comprise nucleic acids that are at least 50% pure.

24. The array of claim 15, wherein the nucleic acids in the different known locations are at least 80% pure.

25. The array of claim 15, the nucleic acids in the different known locations are at least 90% pure.

26. The array of claim 15, wherein said array is produced by a binary synthesis process, said process comprising the steps of:

providing a planar, solid support, said solid support having a plurality of compounds immobilized on a surface thereof, said compounds having protecting groups coupled thereto; deprotecting a first portion of said plurality of compounds on said surface and not a second portion of said plurality of compounds;

reacting said first portion of said plurality of compounds with a first reactant;

deprotecting at least a third portion of said plurality of compounds on said surface, said third portion comprising a fraction of said first portion of said plurality of compounds;

reacting said at least third portion of said plurality of compounds with a second reactant; and

optionally repeating said binary synthesis steps to produce said nucleic acid array.

27. The array of claim 15, wherein the nucleic acids are covalently attached to the support.

28. An array of nucleic acids, the array comprising:

a planar support having at least a first surface; and

a plurality of different nucleic acids attached to the first surface of the solid support at a density exceeding 10,000 different nucleic acids/cm^2, wherein each of the different nucleic acids is attached to the surface of the solid support in a different known location, and has a different determinable sequence.

29. An array of nucleic acids, the an-ay comprising:

a planar support having at least a first surface; and

a plurality of different nucleic acids attached to the first surface of the solid support at a density exceeding 400 different nucleic acids/cm^2, wherein each of the different nucleic acids is attached to the surface of the solid support in a different know location, has a different determinable sequence, wherein the surface and the support are made from different materials.

30. The array of claim 15, wherein the different known locations are square in shape.

31. The array of claim 15, wherein the substrate is glass.

32. The array of claim 15, wherein the substrate is silicon dioxide.

33. The array of claim 15, wherein the substrate is (poly)tetrafluoroethylene, (poly)vinylidenedifluoride, polystyrene or polycarbonate.

34. The method of claim 15, wherein the substrate is optically transparent.

35. The array of claim 15, where in the substrate is functionalized with groups that attach to the plurality of different nucleic acids.

36. The array of claim 1, wherein the plurality of different oligontucleotides have known sequences.

37. The array of claim 15, wherein the plurality of different nucleic acids have known sequences.

38. The array of claim 28, wherein the plurality of different nucleic acids have known sequences.

39. The array of claim 29, wherein the plurality of different nucleic acids have known sequences.

* * * * *

The Sequence of the Human Genome

J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

Dec
hum
for t

[1]Cel
208
Bea
Gen
947
vers
USA
Univ
Ave
Univ
tal,
MD
York
Engl
USA
of T
den
Med
Hav
850
[12]T
Cer
Life
Isra
stit
Por

*T
ma

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (*17*, *18*). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (*19*).

In 1997, Weber and Myers (*20*) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (*21*). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (*22*), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (*23*). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (*24*). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (*25*). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (*26–28*). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (*29*), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (*30*) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

1 Sources of DNA and Sequencing Methods
2 Genome Assembly Strategy and Characterization
3 Gene Prediction and Annotation
4 Genome Structure
5 Genome Evolution
6 A Genome-Wide Examination of Sequence Variations
7 An Overview of the Predicted Protein-Coding Genes in the Human Genome
8 Conclusions

1 Sources of DNA and Sequencing Methods

Summary. This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genome sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (*31*) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (*32*).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was collected, as well as five specimens of semen collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at www.sciencemag.org/cgi/content/291/5507/1304/DC1). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (*33*).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (*34*).

Current sequencing protocols are based on the dideoxy sequencing method (*35*), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (*36*).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (*37*). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (*26*). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

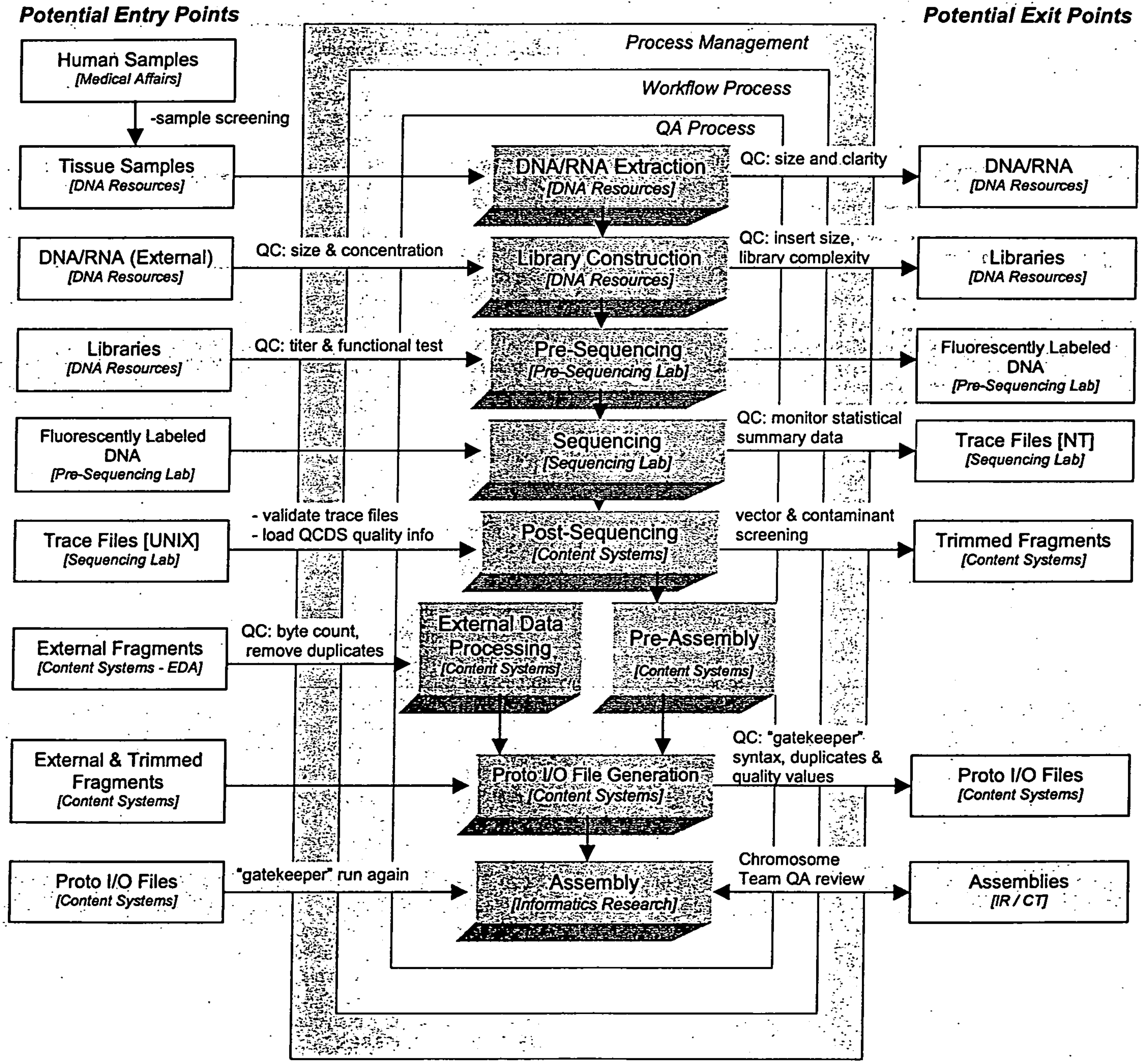

Potential Entry Points
Potential Exit Points
Process Management
Workflow Process
QA Process
Human Samples [Medical Affairs]
-sample screening
Tissue Samples [DNA Resources]
DNA/RNA Extraction [DNA Resources]
QC: size and clarity
DNA/RNA [DNA Resources]
DNA/RNA (External) [DNA Resources]
QC: size & concentration
Library Construction [DNA Resources]
QC: insert size, library complexity
Libraries [DNA Resources]
Libraries [DNA Resources]
QC: titer & functional test
Pre-Sequencing [Pre-Sequencing Lab]
Fluorescently Labeled DNA [Pre-Sequencing Lab]
Fluorescently Labeled DNA [Pre-Sequencing Lab]
Sequencing [Sequencing Lab]
QC: monitor statistical summary data
Trace Files [NT] [Sequencing Lab]
Trace Files [UNIX] [Sequencing Lab]
- validate trace files
- load QCDS quality info
Post-Sequencing [Content Systems]
vector & contaminant screening
Trimmed Fragments [Content Systems]
External Fragments [Content Systems - EDA]
QC: byte count, remove duplicates
External Data Processing [Content Systems]
Pre-Assembly [Content Systems]
External & Trimmed Fragments [Content Systems]
Proto I/O File Generation [Content Systems]
QC: "gatekeeper" syntax, duplicates & quality values
Proto I/O Files [Content Systems]
Proto I/O Files [Content Systems]
"gatekeeper" run again
Assembly [Informatics Research]
Chromosome Team QA review
Assemblies [IR / CT]

Fig. 2. Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange samples and data with both internal and external entities according to defined quality guidelines. Manufacturing pipeline processes, products, quality control measures, and responsible parties are indicated and are described further in the text.

and provide a comparison to the public gen sequence, which was reconstructed largely an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff," and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (PFP) and is primarily derived from BAC clones (*30*). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (*38*), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Seqences division of GenBank (*39*), filtered at 200 bp at 98%; and (iii) the nonredundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (*18*).

2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (*40*). Furthermore, BAC location

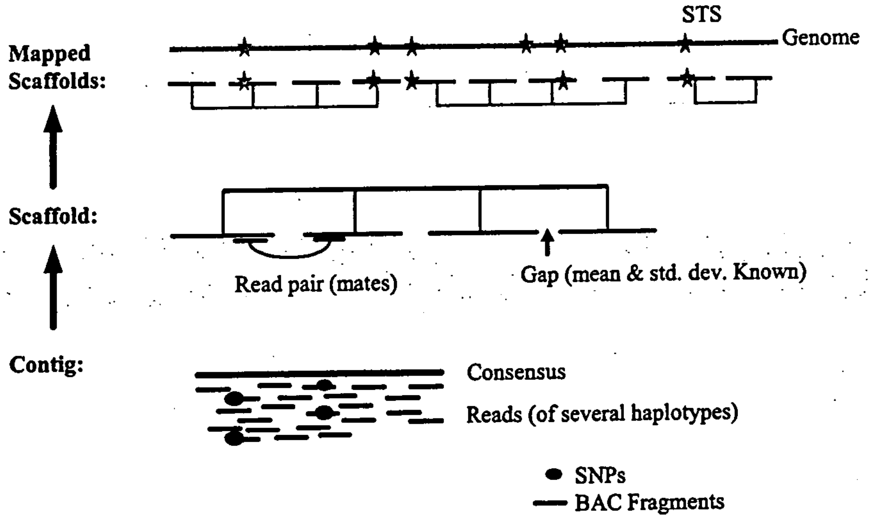


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that at least 2.2% of the BACs contained sequence data that were not part of the given BAC (*41*), possibly as a result of sample-tracking errors (see below). In short, we performed a true, ab initio whole-genome assembly in which we took the expedient of deriving additional sequence coverage, but not mate pairs, assembled bactigs, or genome locality, from some externally generated data.

In the compartmentalized shotgun assembly (CSA), Celera and PFP data were partitioned into the largest possible chromosomal segments or "components" that could be determined with confidence, and then shotgun assembly was applied to each partitioned subset wherein the bactig data were again shredded into faux reads to ensure an independent ab initio assembly of the component. By subsetting the data in this way, the overall computational effort was reduced and the effect of interchromosomal duplications was ameliorated. This also resulted in a reconstruction of the genome that was relatively independent of the whole-genome assembly results so that the two assemblies could be compared for consistency. The quality of the partitioning into components was crucial so that different genome regions were not mixed together. We constructed components from (i) the longest scaffolds of the sequence from each BAC and (ii) assembled scaffolds of data unique to Celera's data set. The BAC assemblies were obtained by a combining assembler that used the bactigs and the 5× Celera data mapped to those bactigs as input. This effort was undertaken as an interim step solely because the more accurate and complete the scaffold for a given sequence stretch, the more accurately one can tile these scaffolds into contiguous components on the basis of sequence overlap and mate-pair information. We further visually inspected and curated the scaffold tiling of the components to further increase its accuracy. For the final CSA assembly, all but the partitioning was ignored, and an independent, ab initio reconstruction of the sequence in each component was obtained by applying our whole-genome assembly algorithm to the partitioned, relevant Celera data and the shredded, faux reads of the partitioned, relevant bactig data.

Table 2. GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

2.3 Whole-genome assembly

The algorithms used for whole-genome assembly (WGA) of the human genome were enhancements to those used to produce the sequence of the *Drosophila* genome reported in detail in (*28*).

The WGA assembler consists of a pipeline composed of five principal stages: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver, respectively. The Screener finds and marks all microsatellite repeats with less than a 6-bp element, and screens out all known interspersed repeat elements, including Alu, Line, and ribosomal DNA. Marked regions get searched for overlaps, whereas screened regions do not get searched, but can be part of an overlap that involves unscreened matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in-10^{17} event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (*42*). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6× simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in 10^{10}, assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than 10^{-7} based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (*43*). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

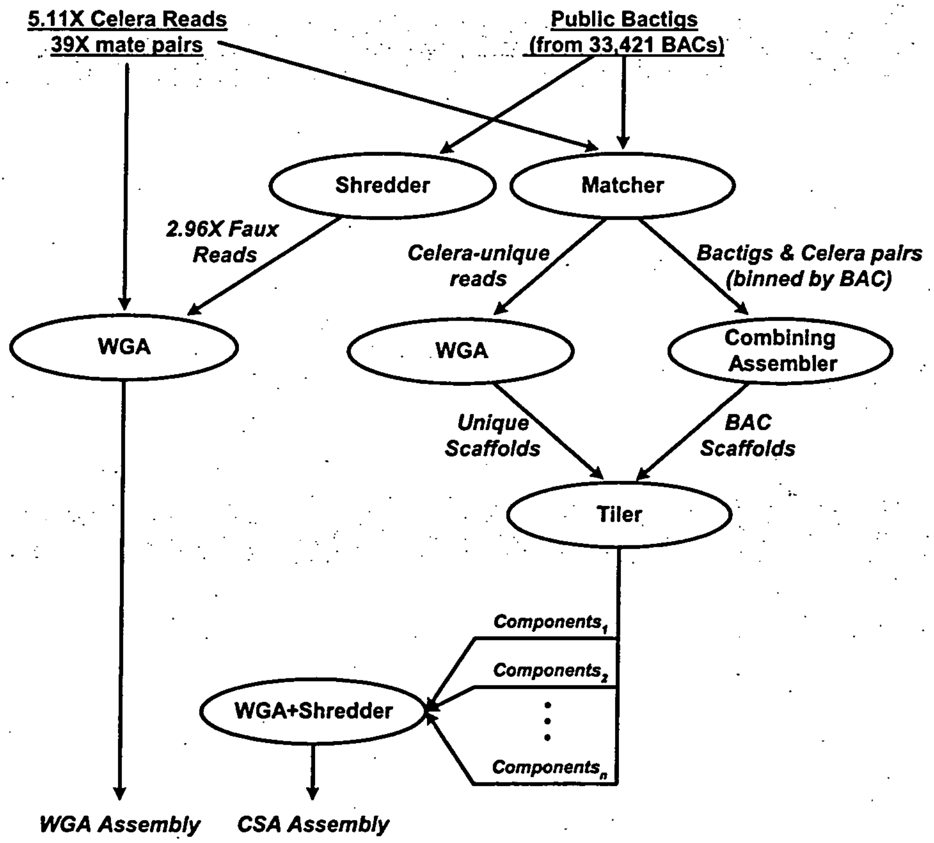


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

some 22, all stones were placed correctly.

The final method of resolving gaps is to fill them with assembled BAC data that cover the gap. We call this external gap "walking." We did not include the very aggressive "Pebbles" substage described in our *Drosophila* work, which made enough mistakes so as to produce repeat reconstructions for long interspersed elements whose quality was only 99.62% correct. We decided that for the human genome it was philosophically better not to introduce a step that was certain to produce less than 99.99% accuracy. The cost was a somewhat larger number of gaps of somewhat larger size.

At the final stage of the assembly process, and also at several intermediate points, a consensus sequence of every contig is produced. Our algorithm is driven by the principle of maximum parsimony, with quality-value–weighted measures for evaluating each base. The net effect is a Bayesian estimate of the correct base to report at each position. Consensus generation uses Celera data whenever it is present. In the event that no Celera data cover a given region, the BAC data sequence is used.

A key element of achieving a WGA of the human genome was to parallelize the Overlapper and the central consensus sequence–constructing subroutines. In addition, memory was a real issue—a straightforward application of the software we had built for *Drosophila* would have required a computer with a 600-gigabyte RAM. By making the Overlapper and Unitigger incremental, we were able to achieve the same computation with a maximum of instantaneous usage of 28 gigabytes of RAM. Moreover, the incremental nature of the first three stages allowed us to continually update the state of this part of the computation as data were delivered and then perform a 7-day run to complete Scaffolding and Repeat Resolution whenever desired. For our assembly operations, the total compute infrastructure consists of 10 four-processor SMPs with 4 gigabytes of memory per cluster (Compaq's ES40, Regatta) and a 16-processor NUMA machine with 64 gigabytes of memory (Compaq's GS160, Wildfire). The total compute for a run of the assembler was roughly 20,000 CPU hours.

The assembly of Celera's data, together with the shredded bactig data, produced a set of scaffolds totaling 2.848 Gbp in span and consisting of 2.586 Gbp of sequence. The chaff, or set of reads not incorporated in the assembly, numbered 11.27 million (26%), which is consistent with our experience for *Drosophila*. More than 84% of the genome was covered by scaffolds >100 kbp long, and these averaged 91% sequence and 9% gaps with a total of 2.297 Gbp of sequence. There were a total of 93,857 gaps among the 1637 scaffolds >100 kbp. The average scaffold size was 1.5 Mbp, the average contig size was 24.06 kbp, and the average gap size was 2.43 kbp, where the distribution of each was essentially exponential. More than 50% of all gaps were less than 500 bp long, >62% of all gaps were less than 1 kbp long, and no gap was >100 kbp long. Similarly, more than 65% of the sequence is in contigs >30 kbp, more than 31% is in contigs >100 kbp, and the largest contig was 1.22 Mbp long. Table 3 gives detailed summary statistics for the structure of this assembly with a direct comparison to the compartmentalized shotgun assembly.

2.4 Compartmentalized shotgun assembly

In addition to the WGA approach, we pursued a localized assembly approach that was intended to subdivide the genome into segments, each of which could be shotgun assembled individually. We expected that this would help in resolution of large interchromosomal duplications and improve the statistics for calculating U-unitigs. The compartmentalized assembly process involved clustering Celera reads and bactigs into large, multiple megabase regions of the genome, and then running the WGA assembler on the Celera data and shredded, faux reads obtained from the bactig data.

The first phase of the CSA strategy was to separate Celera reads into those that matched the BAC contigs for a particular PFP BAC entry, and those that did not match any public data. Such matches must be guaranteed to

Table 3. Scaffold statistics for whole-genome and compartmentalized shotgun assemblies.

	Scaffold size				
	All	>30 kbp	>100 kbp	>500 kbp	>1000 kbp
	Compartmentalized shotgun assembly				
No. of bp in scaffolds (including intrascaffold gaps)	2,905,568,203	2,748,892,430	2,700,489,906	2,489,357,260	2,248,689,128
No. of bp in contigs	2,653,979,733	2,524,251,302	2,491,538,372	2,320,648,201	2,106,521,902
No. of scaffolds	53,591	2,845	1,935	1,060	721
No. of contigs	170,033	112,207	107,199	93,138	82,009
No. of gaps	116,442	109,362	105,264	92,078	81,288
No. of gaps ≤1 kbp	72,091	69,175	67,289	59,915	53,354
Average scaffold size (bp)	54,217	966,219	1,395,602	2,348,450	3,118,848
Average contig size (bp)	15,609	22,496	23,242	24,916	25,686
Average intrascaffold gap size (bp)	2,161	2,054	1,985	1,832	1,749
Largest contig (bp)	1,988,321	1,988,321	1,988,321	1,988,321	1,988,321
% of total contigs	100	95	94	87	79
	Whole-genome assembly				
No. of bp in scaffolds (including intrascaffold gaps)	2,847,890,390	2,574,792,618	2,525,334,447	2,328,535,466	2,140,943,032
No. of bp in contigs	2,586,634,108	2,334,343,339	2,297,678,935	2,143,002,184	1,983,305,432
No. of scaffolds	118,968	2,507	1,637	818	554
No. of contigs	221,036	99,189	95,494	84,641	76,285
No. of gaps	102,068	96,682	93,857	83,823	75,731
No. of gaps ≤1 kbp	62,356	60,343	59,156	54,079	49,592
Average scaffold size (bp)	23,938	1,027,041	1,542,660	2,846,620	3,864,518
Average contig size (bp)	11,702	23,534	24,061	25,319	25,999
Average intrascaffold gap size (bp)	2,560	2,487	2,426	2,213	2,082
Largest contig (bp)	1,224,073	1,224,073	1,224,073	1,224,073	1,224,073
% of total contigs	100	90	89	83	77

properly place a Celera read, so all reads were first masked against a library of common repetitive elements, and only matches of at least 40 bp to unmasked portions of the read constituted a hit. Of Celera's 27.27 million reads, 20.76 million matched a bactig and another 0.62 million reads, which did not have any matches, were nonetheless identified as belonging in the region of the bactig's BAC because their mate matched the bactig. Of the remaining reads, 2.92 million were completely screened out and so could not be matched, but the other 2.97 million reads had unmasked sequence totaling 1.189 Gbp that were not found in the GenBank data set. Because the Celera data are 5.11× redundant, we estimate that 240 Mbp of unique Celera sequence is not in the GenBank data set.

In the next step of the CSA process, a combining assembler took the relevant 5× Celera reads and bactigs for a BAC entry, and produced an assembly of the combined data for that locale. These high-quality sequence reconstructions were a transient result whose utility was simply to provide more reliable information for the purposes of their tiling into sets of overlapping and adjacent scaffold sequences in the next step. In outline, the combining assembler first examines the set of matching Celera reads to determine if there are excessive pileups indicative of unscreened repetitive elements. Wherever these occur, reads in the repeat region whose mates have not been mapped to consistent positions are removed. Then all sets of mate pairs that consistently imply the same relative position of two bactigs are bundled into a link and weighted according to the number of mates in the bundle. A "greedy" strategy then attempts to order the bactigs by selecting bundles of mate-pairs in order of their weight. A selected mate-pair bundle can tie together two formative scaffolds. It is incorporated to form a single scaffold only if it is consistent with the majority of links between contigs of the scaffold. Once scaffolding is complete, gaps are filled by the "Stones" strategy described above for the WGA assembler.

The GenBank data for the Phase 1 and 2 BACs consisted of an average of 19.8 bactigs per BAC of average size 8099 bp. Application of the combining assembler resulted in individual Celera BAC assemblies being put together into an average of 1.83 scaffolds (median of 1 scaffold) consisting of an average of 8.57 contigs of average size 18,973 bp. In addition to defining order and orientation of the sequence fragments, there were 57% fewer gaps in the combined result. For Phase 0 data, the average GenBank entry consisted of 91.52 reads of average length 784 bp. Application of the combining assembler resulted in an average of 54.8 scaffolds consisting of an average of 58.1 contigs of average size 873 bp. Basically, some small amount of [illegible]mbly took place, but not enough Celera data were matched to truly assemble the 0.5× to 1× data set represented by the typical Phase 0 BACs. The combining assembler was also applied to the Phase 3 BACs for SNP identification, confirmation of assembly, and localization of the Celera reads. The phase 0 data suggest that a combined whole-genome shotgun data set and 1× light-shotgun of BACs will not yield good assembly of BAC regions; at least 3× light-shotgun of each BAC is needed.

The 5.89 million Celera fragments not matching the GenBank data were assembled with our whole-genome assembler. The assembly resulted in a set of scaffolds totaling 442 Mbp in span and consisting of 326 Mbp of sequence. More than 20% of the scaffolds were >5 kbp long, and these averaged 63% sequence and 27% gaps with a total of 302 Mbp of sequence. All scaffolds >5 kbp were forwarded along with all scaffolds produced by the combining assembler to the subsequent tiling phase.

At this stage, we typically had one or two scaffolds for every BAC region constituting at least 95% of the relevant sequence, and a collection of disjoint Celera-unique scaffolds. The next step in developing the genome components was to determine the order and overlap tiling of these BAC and Celera-unique scaffolds across the genome. For this, we used Celera's 50-kbp mate-pairs information, and BAC-end pairs (*18*) and sequence tagged site (STS) markers (*44*) to provide long-range guidance and chromosome separation. Given the relatively manageable number of scaffolds, we chose not to produce this tiling in a fully automated manner, but to compute an initial tiling with a good heuristic and then use human curators to resolve discrepancies or missed join opportunities. To this end, we developed a graphical user interface that displayed the graph of tiling overlaps and the evidence for each. A human curator could then explore the implication of mapped STS data, dot-plots of sequence overlap, and a visual display of the mate-pair evidence supporting a given choice. The result of this process was a collection of "components," where each component was a tiled set of BAC and Celera-unique scaffolds that had been curator-approved. The process resulted in 3845 components with an estimated span of 2.922 Gbp.

In order to generate the final CSA, we assembled each component with the WGA algorithm. As was done in the WGA process, the bactig data were shredded into a synthetic 2× shotgun data set in order to give the assembler the freedom to independently assemble the data. By using faux reads rather than bactigs, the assembly algorithm could correct errors in the assembly of bactigs and remove chimeric content in a PFP data entry. C[illegible]r contaminating sequence (from another part of the genome) would not be incorporated into the reassembly of the component because it did not belong there. In effect, the previous steps in the CSA process served only to bring together Celera fragments and PFP data relevant to a large contiguous segment of the genome, wherein we applied the assembler used for WGA to produce an ab initio assembly of the region.

WGA assembly of the components resulted in a set of scaffolds totaling 2.906 Gbp in span and consisting of 2.654 Gbp of sequence. The chaff, or set of reads not incorporated into the assembly, numbered 6.17 million, or 22%. More than 90.0% of the genome was covered by scaffolds spanning >100 kbp long, and these averaged 92.2% sequence and 7.8% gaps with a total of 2.492 Gbp of sequence. There were a total of 105,264 gaps among the 107,199 contigs that belong to the 1940 scaffolds spanning >100 kbp. The average scaffold size was 1.4 Mbp, the average contig size was 23.24 kbp, and the average gap size was 2.0 kbp where each distribution of sizes was exponential. As such, averages tend to be underrepresentative of the majority of the data. Figure 5 shows a histogram of the bases in scaffolds of various size ranges. Consider also that more than 49% of all gaps were <500 bp long, more than 62% of all gaps were <1 kbp, and all gaps are <100 kbp long. Similarly, more than 73% of the sequence is in contigs > 30 kbp, more than 49% is in contigs >100 kbp, and the largest contig was 1.99 Mbp long. Table 3 provides summary statistics for the structure of this assembly with a direct comparison to the WGA assembly.

2.5 Comparison of the WGA and CSA scaffolds

Having obtained two assemblies of the human genome via independent computational processes (WGA and CSA), we compared scaffolds from the two assemblies as another means of investigating their completeness, consistency, and contiguity. From each assembly, a set of reference scaffolds containing at least 1000 fragments (Celera sequencing reads or bactig shreds) was obtained; this amounted to 2218 WGA scaffolds and 1717 CSA scaffolds, for a total of 2.087 Gbp and 2.474 Gbp. The sequence of each reference scaffold was compared to the sequence of all scaffolds from the other assembly with which it shared at least 20 fragments or at least 20% of the fragments of the smaller scaffold. For each such comparison, all matches of at least 200 bp with at most 2% mismatch were tabulated.

From this tabulation, we estimated the amount of unique sequence in each assembly in two ways. The first was to determine the number of bases of each assembly that were not covered by a matching segment in the other assembly. Some 82.5 Mbp of the WGA (3.95%) was not covered by the CSA, whereas 204.5 Mbp (8.26%) of the CSA was not covered by the WGA. This estimate did not require any consistency of the assemblies or any uniqueness of the matching segments. Thus, another analysis was conducted in which matches of less than 1 kbp between a pair of scaffolds were excluded unless they were confirmed by other matches having a consistent order and orientation. This gives some measure of consistent coverage: 1.982 Gbp (95.00%) of the WGA is covered by the CSA, and 2.169 Gbp (87.69%) of the CSA is covered by the WGA by this more stringent measure.

The comparison of WGA to CSA also permitted evaluation of scaffolds for structural inconsistencies. We looked for instances in which a large section of a scaffold from one assembly matched only one scaffold from the other assembly, but failed to match over the full length of the overlap implied by the matching segments. An initial set of candidates was identified automatically, and then each candidate was inspected by hand. From this process, we identified 31 instances in which the assemblies appear to disagree in a nonlocal fashion. These cases are being further evaluated to determine which assembly is in error and why.

In addition, we evaluated local inconsistencies of order or orientation. The following results exclude cases in which one contig in one assembly corresponds to more than one overlapping contig in the other assembly (as long as the order and orientation of the latter agrees with the positions they match in the former). Most of these small rearrangements involved segments on the order of hundreds of base pairs and rarely >1 kbp. We found a total of 295 kbp (0.012%) in the CSA assemblies that were locally inconsistent with the WGA assemblies, whereas 2.108 Mbp (0.11%) in the WGA assembly were inconsistent with the CSA assembly.

The CSA assembly was a few percentage points better in terms of coverage and slightly more consistent than the WGA, because it was in effect performing a few thousand shotgun assemblies of megabase-sized problems, whereas the WGA is performing a shotgun assembly of a gigabase-sized problem. When one considers the increase of two-and-a-half orders of magnitude in problem size, the information loss between the two is remarkably small. Because CSA was logistically easier to deliver and the better of the two results available at the time when downstream analyses needed to be begun, all subsequent analysis was performed on this assembly.

2.6 Mapping scaffolds to the genome

The final step in assembling the genome was to order and orient the scaffolds on the chromosomes. We first grouped scaffolds together on the basis of their order in the components from CSA. These grouped scaffolds were reordered by examining residual mate-pairing data between the scaffolds. We next mapped the scaffold groups onto the chromosome using physical mapping data. This step depends on having reliable high-resolution map information such that each scaffold will overlap multiple markers. There are two genome-wide types of map information available: high-density STS maps and fingerprint maps of BAC clones developed at Washington University (*45*). Among the genome-wide STS maps, GeneMap99 (GM99) has the most markers and therefore was most useful for mapping scaffolds. The two different mapping approaches are complementary to one another. The fingerprint maps should have better local order because they were built by comparison of overlapping BAC clones. On the other hand, GM99 should have a more reliable long-range order, because the framework markers were derived from well-validated genetic maps. Both types of maps were used as a reference for human curation of the components that were the input to the regional assembly, but they did not determine the order of sequences produced by the assembler.

In order to determine the effectiveness of the fingerprint maps and GM99 for mapping scaffolds, we first examined the reliability of these maps by comparison with large scaffolds. Only 1% of the STS markers on the 10 largest scaffolds (those >9 Mbp) were mapped on a different chromosome on GM99. Two percent of the STS markers disagreed in position by more than five framework bins. However, for the fingerprint maps, a 2% chromosome discrepancy was observed, and on average 23.8% of BAC locations in the scaffold sequence disagreed with fingerprint map placement by more than five BACs. When further examining the source of discrepancy, it was found that most of the discrepancy came from 4 of the 10 scaffolds, indicating this there is variation in the quality of either the map or the scaffolds. All four scaffolds were assembled, as well as the other six, as judged by clone coverage analysis, and showed the same low discrepancy rate to GM99, and thus we concluded that the fingerprint map global order in these cases was not reliable. Smaller scaffolds had a higher discordance rate with GM99 (4.21% of STSs were discordant by more than five framework bins), but a lower discordance rate with the fingerprint maps (11% of BACs disagreed with fingerprint maps by more than five BACs). This observation agrees with the clone coverage analysis (*46*) that Celera scaffold construction was better supported by long-range mate pairs in larger scaffolds than in small scaffolds.

We created two orderings of Celera scaffolds on the basis of the markers (BAC or STS) on these maps. Where the order of scaffolds agreed between GM99 and the WashU BAC map, we had a high degree of confidence that that order was correct; these scaffolds were termed "anchor scaffolds." Only scaffolds with a low overall discrepancy rate with both maps were considered anchor scaffolds. Scaffolds in GM99 bins were allowed to permute in their order to match WashU ordering, provided they did not violate their framework orders. Orientation of individual scaffolds was determined by the presence of multiple mapped markers with consistent order. Scaffolds with only one marker have insufficient information to assign orientation. We found 70.1% of the genome in anchored scaffolds, more than 99% of which are also oriented (Table 4). Because GM99 is of lower resolution than the WashU map, a number of scaffolds without STS matches could be ordered relative to the anchored scaffolds because they included sequence from the same or adjacent BACs on the WashU map. On the other hand, because of occasional WashU global ordering discrepancies, a number of scaffolds determined to be "unmappable" on the WashU map could be ordered relative to the anchored scaffolds

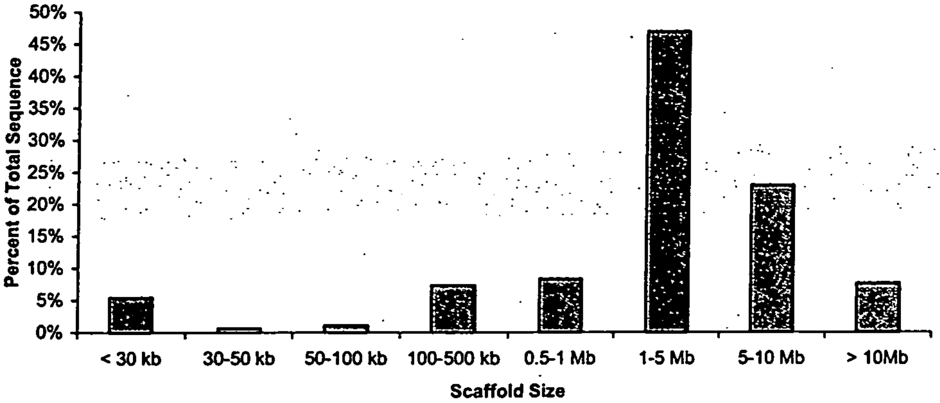


Fig. 5. Distribution of scaffold sizes of the CSA. For each range of scaffold sizes, the percent of total sequence is indicated.

with GM99. These scaffolds were termed "ordered scaffolds." We found that 13.9% of the assembly could be ordered by these additional methods, and thus 84.0% of the genome was ordered unambiguously.

Next, all scaffolds that could be placed, but not ordered, between anchors were assigned to the interval between the anchored scaffolds and were deemed to be "bounded" between them. For example, small scaffolds having STS hits from the same GeneMap bin or hitting the same BAC cannot be ordered relative to each other, but can be assigned a placement boundary relative to other anchored or ordered scaffolds. The remaining scaffolds either had no localization information, conflicting information, or could only be assigned to a generic chromosome location. Using the above approaches, ~98% of the genome was anchored, ordered, or bounded.

Finally, we assigned a location for each scaffold placed on the chromosome by spreading out the scaffolds per chromosome. We assumed that the remaining unmapped scaffolds, constituting 2% of the genome, were distributed evenly across the genome. By dividing the sum of unmapped scaffold lengths with the sum of the number of mapped scaffolds, we arrived at an estimate of interscaffold gap of 1483 bp. This gap was used to separate all the scaffolds on each chromosome and to assign an offset in the chromosome.

During the scaffold-mapping effort, we encountered many problems that resulted in additional quality assessment and validation analysis. At least 978 (3% of 33,173) BACs were believed to have sequence data from more than one location in the genome (*47*). This is consistent with the bactig chimerism analysis reported above in the Assembly Strategies section. These BACs could not be assigned to unique positions within the CSA assembly and thus could not be used for ordering scaffolds. Likewise, it was not always possible to assign STSs to unique locations in the assembly because of genome duplications, repetitive elements, and pseudogenes.

Because of the time required for an exhaustive search for a perfect overlap, CSA generated 21,607 intrascaffold gaps where the mate-pair data suggested that the contigs should overlap, but no overlap was found. These gaps were defined as a fixed 50 bp in length and make up 18.6% of the total 116,442 gaps in the CSA assembly.

We chose not to use the order of exons implied in cDNA or EST data as a way of ordering scaffolds. The rationale for not using this data was that doing so would have biased certain regions of the assembly by rearranging scaffolds to fit the transcript data and made validation of both the assembly and gene definition processes more difficult.

Assembly and validation analysis

We analyzed the assembly of the genome from the perspectives of completeness (amount of coverage of the genome) and correctness (the structural accuracy of the order and orientation and the consensus sequence of the assembly).

Completeness. Completeness is defined as the percentage of the euchromatic sequence represented in the assembly. This cannot be known with absolute certainty until the euchromatin sequence has been completed. However, it is possible to estimate completeness on the basis of (i) the estimated sizes of intrascaffold gaps; (ii) coverage of the two published chromosomes, 21 and 22 (*48*, *49*); and (iii) analysis of the percentage of an independent set of random sequences (STS markers) contained in the assembly. The whole-genome libraries contain heterochromatic sequence and, although no attempt has been made to assemble it, there may be instances of unique sequence embedded in regions of heterochromatin as were observed in *Drosophila* (*50*, *51*).

The sequences of human chromosomes 21 and 22 have been completed to high quality and published (*48*, *49*). Although this sequence served as input to the assembler, the finished sequence was shredded into a shotgun data set so that the assembler had the opportunity to assemble it differently from the original sequence in the case of structural polymorphisms or assembly errors in the BAC data. In particular, the assembler must be able to resolve repetitive elements at the scale of components (generally multimegabase in size), and so this comparison reveals the level to which the assembler resolves repeats. In certain areas, the assembly structure differs from the published versions of chromosomes 21 and 22 (see below). The consequence of the flexibility to assemble "finished" sequence differently on the basis of Celera data resulted in an assembly with more segments than the chromosome 21 and 22 sequences. We examined the reasons why there are more gaps in the Celera sequence than in chromosomes 21 and 22 and expect that they may be typical of gaps in other regions of the genome. In the Celera assembly, there are 25 scaffolds, each containing at least 10 kb of sequence, that collectively span 94.3% of chromosome 21. Sixty-two scaffolds span 95.7% of chromosome 22. The total length of the gaps remaining in the Celera assembly for these two chromosomes is 3.4 Mbp. These gap sequences were analyzed by RepeatMasker and by searching against the entire genome assembly (*52*). About 50% of the gap sequence consisted of common repetitive elements identified by RepeatMasker; more than half of the remainder was lower copy number repeat elements.

A more global way of assessing completeness is to measure the content of an independent set of sequence data in the assembly. We compared 48,938 STS markers from Genemap99 (*51*) to the scaffolds. Because these markers were not used in the assembly processes, they provided a truly independent measure of completeness. ePCR (*53*) and BLAST (*54*) were used to locate STSs on the assembled genome. We found 44,524 (91%) of the STSs in the mapped genome. An additional 2648 markers (5.4%) were found by searching the unassembled data or "chaff." We identified 1283 STS markers (2.6%) not found in either Celera sequence or BAC data as of September 2000, raising the possibility that these markers may not be of human origin. If that were the case, the Celera assembled sequence would represent 93.4% of the human genome and the unassembled data 5.5%, for a total of 98.9% coverage. Similarly, we compared CSA against 36,678 TNG radiation hybrid markers (*55a*) using the same method. We found that 32,371 markers (88%) were located in the mapped CSA scaffolds, with 2055 markers (5.6%) found in the remainder. This gave a 94% coverage of the genome through another genome-wide survey.

Correctness. Correctness is defined as the structural and sequence accuracy of the assembly. Because the source sequences for the Celera data and the GenBank data are from different individuals, we could not directly compare the consensus sequence of the as-

Table 4. Summary of scaffold mapping. Scaffolds were mapped to the genome with different levels of confidence (anchored scaffolds have the highest confidence; unmapped scaffolds have the lowest). Anchored scaffolds were consistently ordered by the WashU BAC map and GM99. Ordered scaffolds were consistently ordered by at least one of the following: the WashU BAC map, GM99, or component tiling path. Bounded scaffolds had order conflicts between at least two of the external maps, but their placements were adjacent to a neighboring anchored or ordered scaffold. Unmapped scaffolds had, at most, a chromosome assignment. The scaffold subcategories are given below each category.

Mapped scaffold category	Number	Length (bp)	% Total length
Anchored	1,526	1,860,676,676	70
Oriented	1,246	1,852,088,645	70
Unoriented	280	8,588,031	0.3
Ordered	2,001	369,235,857	14
Oriented	839	329,633,166	12
Unoriented	1,162	39,602,691	2
Bounded	38,241	368,753,463	14
Oriented	7,453	274,536,424	10
Unoriented	30,788	94,217,039	4
Unmapped	11,823	55,313,737	2
Known chromosome	281	2,505,844	0.1
Unknown chromosome	11,542	52,807,893	2

sembly against other finished sequence for determining sequencing accuracy at the nucleotide level, although this has been done for identifying polymorphisms as described in Section 6. The accuracy of the consensus sequence is at least 99.96% on the basis of a statistical estimate derived from the quality values of the underlying reads.

The structural consistency of the assembly can be measured by mate-pair analysis. In a correct assembly, every mated pair of sequencing reads should be located on the consensus sequence with the correct separation and orientation between the pairs. A pair is termed "valid" when the reads are in the correct orientation and the distance between them is within the mean ± 3 standard deviations of the distribution of insert sizes of the library from which the pair was sampled. A pair is termed "misoriented" when the reads are not correctly oriented, and is termed "misseparated" when the distance between the reads is not in the correct range but the reads are correctly oriented. The mean ± the standard deviation of each library used by the assembler was determined as described above. To validate these, we examined all reads mapped to the finished sequence of chromosome 21 (*48*) and determined how many incorrect mate pairs there were as a result of laboratory tracking errors and chimerism (two different segments of the genome cloned into the same plasmid), and how tight the distribution of insert sizes was for those that were correct (Table 5). The standard deviations for all Celera libraries were quite small, less than 15% of the insert length, with the exception of a few 50-kbp libraries. The 2- and 10-kbp libraries contained less than 2% invalid mate pairs, whereas the 50-kbp libraries were somewhat higher (~10%). Thus, although the mate-pair information was not perfect, its accuracy was such that measuring valid, misoriented, and misseparated pairs with respect to a given assembly was deemed to be a reliable instrument for validation purposes, especially when several mate pairs confirm or deny an ordering.

The clone coverage of the genome was 39×, meaning that any given base pair was, on average, contained in 39 clones or, equivalently, spanned by 39 mate-paired reads. Areas of low clone coverage or areas with a high proportion of invalid mate pairs would indicate potential assembly problems. We computed the coverage of each base in the assembly by valid mate pairs (Table 6). In summary, for scaffolds >30 kbp in length, less than 1% of the Celera assembly was in regions of less than 3× clone coverage. Thus, more than 99% of the assembly, including order and orientation, is strongly supported by this measure alone.

We examined the locations and number of all misoriented and misseparated mates. In addition to doing this analysis on the CSA assembly (as of 1 October 2000), we also performed a study of the PFP assembly as of 5 September 2000 (*30*, *55b*). In this latter case, Celera mate pairs had to be mapped to the PFP assembly. To avoid mapping errors due to high-fidelity repeats, the only pairs mapped were those for which both reads matched at only one location with less than 6% differences. A threshold was set such that sets of five or more simultaneously invalid mate pairs indicated a potential breakpoint, where the construction of the two assemblies differed. The graphic comparison of the CSA chromosome 21 assembly with the published sequence (Fig. 6A) serves as a validation of this methodology. Blue tick marks in the panels indicate breakpoints. There were a similar (small) number of breakpoints on both chromosome sequences. The exception was 12 sets of scaffolds in the Celera assembly (a total of 3% of the chromosome length in 212 single-contig scaffolds) that were mapped to the wrong positions because they were too small to be mapped reliably. Figures 6 and 7 and Table 6 illustrate the mate-pair differences and breakpoints between the two assemblies. There was a higher percentage of misoriented and misseparated mate pairs in the large-insert libraries (50 kbp and BAC ends) than in the small-insert libraries in both assemblies (Table 6). The large-insert libraries are more likely to identify discrepancies simply because they span a larger segment of the genome. The graphic comparison between the two assemblies for chromosome 8 (Fig. 6, B and C) shows that there are many

Table 5. Mate-pair validation. Celera fragment sequences were mapped to the published sequence of chromosome 21. Each mate pair uniquely mapped was evaluated for correct orientation and placement (number of mate pairs tested). If the two mates had incorrect relative orientation or placement, they were considered invalid (number of invalid mate pairs).

		Chromosome 21						Genome		
Library type	Library no.	Mean insert size (bp)	SD (bp)	SD/ mean (%)	No. of mate pairs tested	No. of invalid mate pairs	% invalid	Mean insert size (bp)	SD (bp)	SD/ mean (%)
2 kbp	1	2,081	106	5.1	3,642	38	1.0	2,082	90	4.3
	2	1,913	152	7.9	28,029	413	1.5	1,923	118	6.1
	3	2,166	175	8.1	4,405	57	1.3	2,162	158	7.3
10 kbp	4	11,385	851	7.5	4,319	80	1.9	11,370	696	6.1
	5	14,523	1,875	12.9	7,355	156	2.1	14,142	1,402	9.9
	6	9,635	1,035	10.7	5,573	109	2.0	9,606	934	9.7
	7	10,223	928	9.1	34,079	399	1.2	10,190	777	7.6
50 kbp	8	64,888	2,747	4.2	16	1	6.3	65,500	5,504	8.4
	9	53,410	5,834	10.9	914	170	18.6	53,311	5,546	10.4
	10	52,034	7,312	14.1	5,871	569	9.7	51,498	6,588	12.8
	11	52,282	7,454	14.3	2,629	213	8.1	52,282	7,454	14.3
	12	46,616	7,378	15.8	2,153	215	10.0	45,418	9,068	20.0
	13	55,788	10,099	18.1	2,244	249	11.1	53,062	10,893	20.5
	14	39,894	5,019	12.6	199	7	3.5	36,838	9,988	27.1
BES	15	48,931	9,813	20.1	144	10	6.9	47,845	4,774	10.0
	16	48,130	4,232	8.8	195	14	7.2	47,924	4,581	9.6
	17	106,027	27,778	26.2	330	16	4.8	152,000	26,600	17.5
	18	160,575	54,973	34.2	155	8	5.2	161,750	27,000	16.7
	19	164,155	19,453	11.9	642	44	6.9	176,500	19,500	11.05
Sum					102,894	2,768 (mean = 2.7)	2.7			

more breakpoints for the PFP assembly than for the Celera assembly. Figure 7 shows the breakpoint map (blue tick marks) for both assemblies of each chromosome in a side-by-side fashion. The order and orientation of Celera's assembly shows substantially fewer breakpoints except on the two finished chromosomes. Figure 7 also depicts large gaps (>10 kbp) in both assemblies as red tick marks. In the CSA assembly, the size of all gaps have been estimated on the basis of the mate-pair data. Breakpoints can be caused by structural polymorphisms, because the two assemblies were derived from different human genomes. They also reflect the unfinished nature of both genome assemblies.

3 Gene Prediction and Annotation

Summary. To enumerate the gene inventory, we developed an integrated, evidence-based approach named Otto. The evidence used to increase the likelihood of identifying genes includes regions conserved between the mouse and human genomes, similarity to ESTs or other mRNA-derived data, or similarity to other proteins. A comparison of Otto (combined Otto-RefSeq and Otto homology) with Genscan, a standard gene-prediction algorithm, showed greater sensitivity (0.78 versus 0.50) and specificity (0.93 versus 0.63) of Otto in the ability to define gene structure. Otto-predicted genes were complemented with a set of genes from three gene-prediction programs that exhibited weaker, but still significant, evidence that they may be expressed. Conservative criteria, requiring at least two lines of evidence, were used to define a set of 26,383 genes with good confidence that were used for more detailed analysis presented in the subsequent sections. Extensive manual curation to establish precise characterization of gene structure will be necessary to improve the results from this initial computational approach.

3.1 Automated gene annotation

A gene is a locus of cotranscribed exons. A single gene may give rise to multiple transcripts, and thus multiple distinct proteins with multiple functions, by means of alternative splicing and alternative transcription initiation and termination sites. Our cells are able to discern within the billions of base pairs of the genomic DNA the signals for initiating transcription and for splicing together exons separated by a few or hundreds of thousands of base pairs. The first step in characterizing the genome is to define the structure of each gene and each transcription unit.

The number of protein-coding genes in mammals has been controversial from the outset. Initial estimates based on reassociation data placed it between 30,000 to 40,000, whereas later estimates from the brain were >100,000 (*56*). More recent data from both the corporate and public sectors, based on extrapolations from EST, CpG island, and transcript density–based extrapolations, have not reduced this variance. The highest recent number of 142,634 genes emanates from a report from Incyte Pharmaceuticals, and is based on a combination of EST data and the association of ESTs with CpG islands (*57*). In stark contrast are three quite different, and much lower estimates: one of ~35,000 genes derived with genome-wide EST data and sampling procedures in conjunction with chromosome 22 data (*58*); another of 28,000 to 34,000 genes derived with a comparative methodology involving sequence conservation between humans and the puffer fish *Tetraodon nigroviridis* (*59*); and a figure of 35,000 genes, which was derived simply by extrapolating from the density of 770 known and predicted genes in the 67 Mbp of chromosomes 21 and 22, to the approximately 3-Gbp euchromatic genome.

The problem of computational identification of transcriptional units in genomic DNA sequence can be divided into two phases. The first is to partition the sequence into segments that are likely to correspond to individual genes. This is not trivial and is a weakness of most de novo gene-finding algorithms. It is also critical to determining the number of genes in the human gene inventory. The second challenge is to construct a gene model that reflects the probable structure of the transcript(s) encoded in the region. This can be done with reasonable accuracy when a full-length cDNA has been sequenced or a highly homologous protein sequence is known. De novo gene prediction, although less accurate, is the only way to find genes that are not represented by homologous proteins or ESTs. The following section describes the methods we have developed to address these problems for the prediction of protein-coding genes.

We have developed a rule-based expert system, called Otto, to identify and characterize genes in the human genome (*60*). Otto attempts to simulate in software the process that a human annotator uses to identify a gene and refine its structure. In the process of annotating a region of the genome, a human curator examines the evidence provided by the computational pipeline (described below) and examines how various types of evidence relate to one another. A curator puts different levels of confidence in different types of evidence and looks for certain patterns of evidence to support gene annotation. For example, a curator may examine homology to a number of ESTs and evaluate whether or not they can be connected into a longer, virtual mRNA. The curator would also evaluate the strength of the similarity and the contiguity of the match, in essence asking whether any ESTs cross splice-junctions and whether the edges of putative exons have consensus splice sites. This kind of manual annotation process was used to annotate the *Drosophila* genome.

The Otto system can promote observed evidence to a gene annotation in one of two ways. First, if the evidence includes a high-quality match to the sequence of a known gene [here defined as a human gene represented in a curated subset of the RefSeq database (*61*)], then Otto can promote this to a gene annotation. In the second method, Otto evaluates a broad spectrum of evidence and determines if this evidence is adequate to support promotion to a gene annotation. These processes are described below.

Initially, gene boundaries are predicted on the basis of examination of sets of overlapping protein and EST matches generated by a computational pipeline (*62*). This pipeline searches the scaffold sequences against protein, EST, and genome-sequence databases to define regions of sequence similarity and runs three de novo gene-prediction programs.

To identify likely gene boundaries, regions of the genome were partitioned by Otto on the basis of sequence matches identified by BLAST. Each of the database sequences matched in the region under analysis was compared by an algorithm that takes into account both coordinates of the matching sequence, as well as the sequence type (e.g., protein, EST, and so forth). The results were used to group the matches into bins of related sequences that may define a gene and identify

Table 6. Genome-wide mate pair analysis of compartmentalized shotgun (CSA) and PFP assemblies.*

Genome library	CSA			PFP		
	% valid	% mis-oriented	% mis-separated†	% valid	% mis-oriented	% mis-separated†
2 kbp	98.5	0.6	1.0	95.7	2.0	2.3
10 kbp	96.7	1.0	2.3	81.9	9.6	8.6
50 kbp	93.9	4.5	1.5	64.2	22.3	13.5
BES	94.1	2.1	3.8	62.0	19.3	18.8
Mean	97.4	1.0	1.6	87.3	6.8	5.9

*Data for individual chromosomes can be found in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1. †Mates are misseparated if their distance is >3 SD from the mean library size.

gene boundaries. During this process, multiple hits to the same region were collapsed to a coherent set of data by tracking the coverage of a region. For example, if a group of bases was represented by multiple overlapping ESTs, the union of these regions matched by the set of ESTs on the scaffold was marked as being supported by EST evidence. This resulted in a series of "gene bins," each of which was believed to contain a single gene. One weakness of this initial implementation of the algorithm was in predicting gene boundaries in regions of tandemly duplicated genes. Gene clusters frequently resulted in homologous neighboring genes being joined together, resulting in an annotation that artificially concatenated these gene models.

Next, known genes (those with exact matches of a full-length cDNA sequence to the genome) were identified, and the region corresponding to the cDNA was annotated as a predicted transcript. A subset of the curated human gene set RefSeq from the National Center for Biotechnology Information (NCBI) was included as a data set searched in the computational pipeline. If a RefSeq transcript matched the genome assembly for at least 50% of its length at >92% identity, then the SIM4 (*63*) alignment of the RefSeq transcript to the region of the genome under analysis was promoted to the status of an Otto annotation. Because the genome sequence has gaps and sequence errors such as frameshifts, it was not always possible to predict a transcript that agrees precisely with the experimentally determined cDNA sequence. A total of 6538 genes in our inventory were identified and transcripts predicted in this way.

Regions that have a substantial amount of sequence similarity, but do not match known genes, were analyzed by that part of the Otto system that uses the sequence similarity information to predict a transcript. Here, Otto

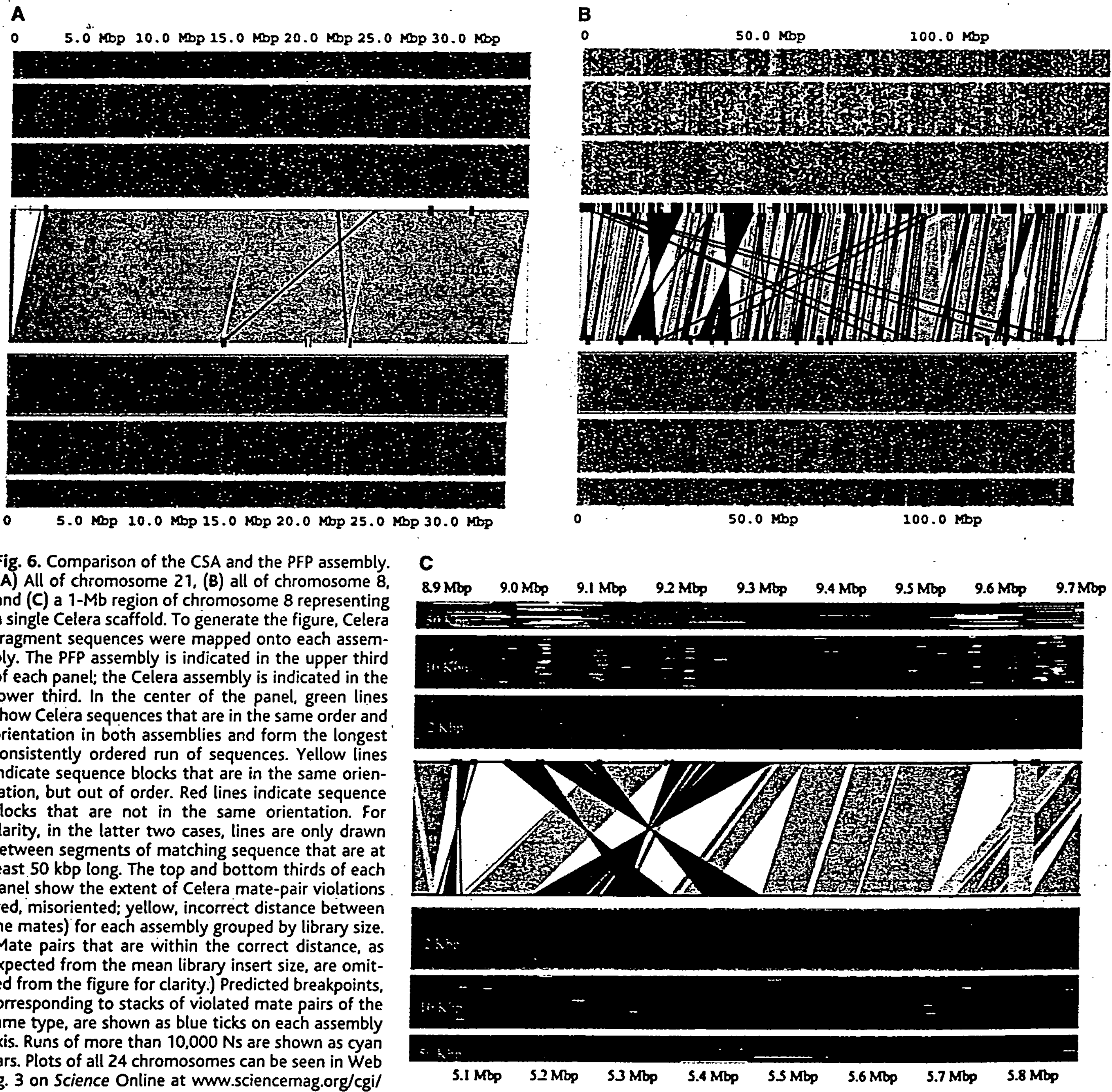


Fig. 6. Comparison of the CSA and the PFP assembly. (**A**) All of chromosome 21, (**B**) all of chromosome 8, and (**C**) a 1-Mb region of chromosome 8 representing a single Celera scaffold. To generate the figure, Celera fragment sequences were mapped onto each assembly. The PFP assembly is indicated in the upper third of each panel; the Celera assembly is indicated in the lower third. In the center of the panel, green lines show Celera sequences that are in the same order and orientation in both assemblies and form the longest consistently ordered run of sequences. Yellow lines indicate sequence blocks that are in the same orientation, but out of order. Red lines indicate sequence blocks that are not in the same orientation. For clarity, in the latter two cases, lines are only drawn between segments of matching sequence that are at least 50 kbp long. The top and bottom thirds of each panel show the extent of Celera mate-pair violations (red, misoriented; yellow, incorrect distance between the mates) for each assembly grouped by library size. (Mate pairs that are within the correct distance, as expected from the mean library insert size, are omitted from the figure for clarity.) Predicted breakpoints, corresponding to stacks of violated mate pairs of the same type, are shown as blue ticks on each assembly axis. Runs of more than 10,000 Ns are shown as cyan bars. Plots of all 24 chromosomes can be seen in Web fig. 3 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1.

evaluates evidence generated by the computational pipeline, corresponding to conservation between mouse and human genomic DNA, similarity to human transcripts (ESTs and cDNAs), similarity to rodent transcripts (ESTs and cDNAs), and similarity of the translation of human genomic DNA to known proteins to predict potential genes in the human genome. The sequence from the region of genomic DNA contained in a gene bin was extracted, and the subsequences supported by any homology evidence were marked (plus 100

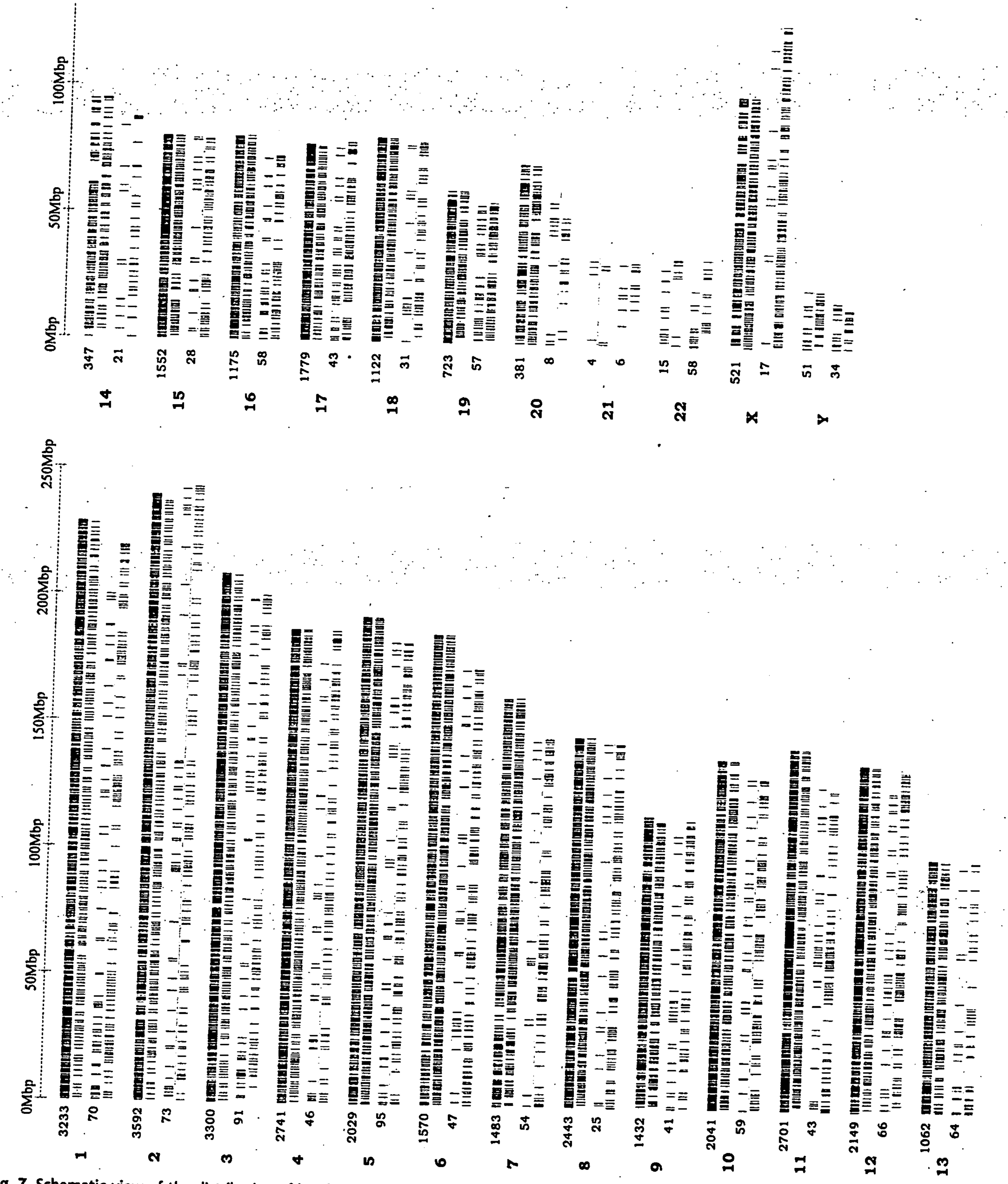


Fig. 7. Schematic view of the distribution of breakpoints and large gaps on all chromosomes. For each chromosome, the upper pair of lines represent the PFP assembly, and the lower pair of lines represent Celera's assembly. Blue tick marks represent breakpoints, whereas red tick marks represent a gap of larger than 10,000 bp. The number of breakpoints per chromosome is indicated in black, and the chromosome numbers in red.

bases flanking these regions). The other bases in the region, those not covered by any homology evidence, were replaced by N's. This sequence segment, with high confidence regions represented by the consensus genomic sequence and the remainder represented by N's, was then evaluated by Genscan to see if a consistent gene model could be generated. This procedure simplified the gene-prediction task by first establishing the boundary for the gene (not a strength of most gene-finding algorithms), and by eliminating regions with no supporting evidence. If Genscan returned a plausible gene model, it was further evaluated before being promoted to an "Otto" annotation. The final Genscan predictions were often quite different from the prediction that Genscan returned on the same region of native genomic sequence. A weakness of using Genscan to refine the gene model is the loss of valid, small exons from the final annotation.

The next step in defining gene structures based on sequence similarity was to compare each predicted transcript with the homology-based evidence that was used in previous steps to evaluate the depth of evidence for each exon in the prediction. Internal exons were considered to be supported if they were covered by homology evidence to within ±10 bases of their edges. For first and last exons, the internal edge was required to be within 10 bases, but the external edge was allowed greater latitude to allow for 5′ and 3′ untranslated regions (UTRs). To be retained, a prediction for a multi-exon gene must have evidence such that the total number of "hits," as defined above, divided by the number of exons in the prediction must be >0.66 or must correspond to a RefSeq sequence. A single-exon gene must be covered by at least three supporting hits (±10 bases on each side), and these must cover the complete predicted open reading frame. For a single-exon gene, we also required that the Genscan prediction include both a start and a stop codon. Gene models that did not meet these criteria were disregarded, and those that passed were promoted to Otto predictions. Homology-based Otto predictions do not contain 3′ and 5′ untranslated sequence. Although three de novo gene-finding programs [GRAIL, Genscan, and FgenesH (*63*)] were run as part of the computational analysis, the results of these programs were not directly used in making the Otto predictions. Otto predicted 11,226 additional genes by means of sequence similarity.

Table 7. Sensitivity and specificity of Otto and Genscan. Sensitivity and specificity were calculated by first aligning the prediction to the published RefSeq transcript, tallying the number (*N*) of uniquely aligned RefSeq bases. Sensitivity is the ratio of *N* to the length of the published RefSeq transcript. Specificity is the ratio of *N* to the length of the prediction. All differences are significant (Tukey HSD; $P < 0.001$).

Method	Sensitivity	Specificity
Otto (RefSeq only)*	0.939	0.973
Otto (homology)†	0.604	0.884
Genscan	0.501	0.633

*Refers to those annotations produced by Otto using only the Sim4-polished RefSeq alignment rather than an evidence-based Genscan prediction. †Refers to those annotations produced by supplying all available evidence to Genscan.

3.2 Otto validation

To validate the Otto homology-based process and the method that Otto uses to define the structures of known genes, we compared transcripts predicted by Otto with their corresponding (and presumably correct) transcript from a set of 4512 RefSeq transcripts for which there was a unique SIM4 alignment (Table 7). In order to evaluate the relative performance of Otto and Genscan, we made three comparisons. The first involved a determination of the accuracy of gene models predicted by Otto with only homology data other than the corresponding RefSeq sequence (Otto homology in Table 7). We measured the sensitivity (correctly predicted bases divided by the total length of the cDNA) and specificity (correctly predicted bases divided by the sum of the correctly and incorrectly predicted bases). Second, we examined the sensitivity and specificity of the Otto predictions that were made solely with the RefSeq sequence, which is the process that Otto uses to annotate known genes (Otto-RefSeq). And third, we determined the accuracy of the Genscan predictions corresponding to these RefSeq sequences. As expected, the alignment method (Otto-RefSeq) was the most accurate, and Otto-homology performed better than Genscan by both criteria. Thus, 6.1% of true RefSeq nucleotides were not represented in the Otto-refseq annotations and 2.7% of the nucleotides in the Otto-RefSeq transcripts were not contained in the original RefSeq transcripts. The discrepancies could come from legitimate differences between the Celera assembly and the RefSeq transcript due to polymorphisms, incomplete or incorrect data in the Celera assembly, errors introduced by Sim4 during the alignment process, or the presence of alternatively spliced forms in the data set used for the comparisons.

Because Otto uses an evidence-based approach to reconstruct genes, the absence of experimental evidence for intervening exons may inadvertantly result in a set of exons that cannot be spliced together to give rise to a transcript. In such cases, Otto may "split genes" when in fact all the evidence should be combined into a single transcript. We also examined the tendency of these methods to incorrectly split gene predictions. These trends are shown in Fig. 8. Both RefSeq and homology-based predictions by Otto split known genes into fewer segments than Genscan alone.

3.3 Gene number

Recognizing that the Otto system is quite conservative, we used a different gene-prediction strategy in regions where the homology evidence was less strong. Here the results of de novo gene predictions were used. For these genes, we insisted that a predicted transcript have at least two of the following types of evidence to be included in the gene set for further analysis: protein, human EST, rodent EST, or mouse genome fragment matches. This final class of predicted genes is a subset of the predictions made by the three gene-finding programs that were used in the computational pipeline. For these, there was not sufficient sequence similarity information for Otto to attempt to predict a gene structure. The three de novo gene-finding programs resulted in about 155,695 predictions, of which ~76,410 were nonredundant (nonoverlapping with one another). Of these, 57,935 did not overlap known genes or predictions made by Otto. Only 21,350 of the gene predictions that did not overlap Otto predictions were partially supported by at least one type of sequence similarity evidence, and 8619 were partially supported by two types of evidence (Table 8).

The sum of this number (21,350) and the number of Otto annotations (17,764), 39,114, is near the upper limit for the human gene complement. As seen in Table 8, if the requirement for other supporting evidence is made more stringent, this number drops rapidly so that demanding two types of evidence reduces the total gene number to 26,383 and demanding three types reduces it to ~23,000. Requiring that a prediction be supported by all four categories of evidence is too stringent because it would eliminate genes that encode novel proteins (members of currently undescribed protein families). No correction for pseudogenes has been made at this point in the analysis.

In a further attempt to identify genes that were not found by the autoannotation process or any of the de novo gene finders, we examined regions outside of gene predictions that were similar to the EST sequence, and where the EST matched the genomic sequence across a splice junction. After correcting for potential 3′ UTRs of predicted genes, about 2500 such regions remained. Addition of a requirement for at least one of the following evidence types—homology to mouse genomic sequence fragments, rodent ESTs, or cDNAs—or similarity to a known protein reduced this number to 1010. Adding this to the numbers from the previous paragraph would give us estimates of about 40,000, 27,000, and 24,000 potential genes in the human genome, depending on the stringency of evidence considered. Table 8 illustrates the number of genes and presents the degree of confidence based on the supporting evidence. Transcripts encoded by a set of 26,383 genes were assembled for further analysis. This set includes the 6538 genes predicted by Otto on the basis of matches to known genes, 11,226 transcripts predicted by Otto based on homology evidence, and 8619 from the subset of transcripts from de novo gene-prediction programs that have two types of supporting evidence. The 26,383 genes are illustrated along chromosome diagrams in Fig. 1. These are a very preliminary set of annotations and are subject to all the limitations of an automated process. Considerable refinement is still necessary to improve the accuracy of these transcript predictions. All the predictions and descriptions of genes and the associated evidence that we present are the product of completely computational processes, not expert curation. We have attempted to enumerate the genes in the human genome in such a way that we have different levels of confidence based on the amount of supporting evidence: known genes, genes with good protein or EST homology evidence, and de novo gene predictions confirmed by modest homology evidence.

3.4 Features of human gene transcripts

We estimate the average span for a "typical" gene in the human DNA sequence to be about 27,894 bases. This is based on the average span covered by RefSeq transcripts, used because it represents our highest confidence set.

The set of transcripts promoted to gene annotations varies in a number of ways. As can be seen from Table 8 and Fig. 9, transcripts predicted by Otto tend to be longer, having on average about 7.8 exons, whereas those promoted from gene-prediction programs average about 3.7 exons. The largest number of exons that we have identified in a transcript is 234 in the titin mRNA. Table 8 compares the amounts of evidence that support the Otto and other predicted transcripts. For example, one can see that a typical Otto transcript has 6.99 of its 7.81 exons supported by protein homology evidence. As would be expected, the Otto transcripts generally have more support than do transcripts predicted by the de novo methods.

4 Genome Structure

Summary. This section describes several of the noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set. These include an analysis of G+C content and gene density in the context of cytogenetic maps of the genome, an enumerative analysis of CpG islands, and a brief description of the genome-wide repetitive elements.

4.1 Cytogenetic maps

Perhaps the most obvious, and certainly the most visible, element of the structure of the genome is the banding pattern produced by Giemsa stain. Chromosomal banding studies have revealed that about 17% to 20% of the human chromosome complement consists of C-bands, or constitutive heterochromatin (*64*). Much of this heterochromatin is highly polymorphic and consists of different families of alpha satellite DNAs with various higher order repeat structures (*65*). Many chromosomes have complex inter- and intrachromosomal duplications present in pericentromeric regions (*66*). About 5% of the sequence reads were identified as alpha satellite sequences; these were not included in the assembly.

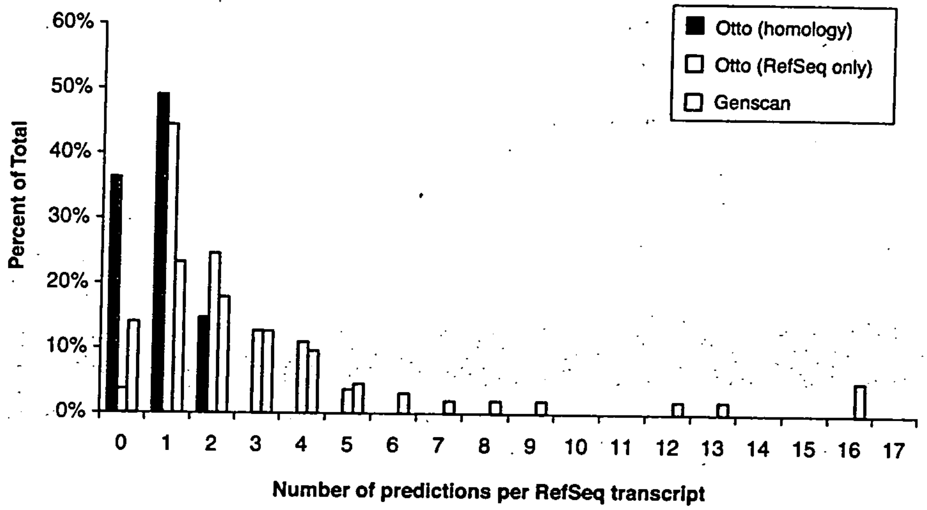


Fig. 8. Analysis of split genes resulting from different annotation methods. A set of 4512 Sim4-based alignments of RefSeq transcripts to the genomic assembly were chosen (see the text for criteria), and the numbers of overlapping Genscan, Otto (RefSeq only) annotations based solely on Sim4-polished RefSeq alignments, and Otto (homology) annotations (annotations produced by supplying all available evidence to Genscan) were tallied. These data show the degree to which multiple Genscan predictions and/or Otto annotations were associated with a single RefSeq transcript. The zero class for the Otto-homology predictions shown here indicates that the Otto-homology calls were made without recourse to the RefSeq transcript, and thus no Otto call was made because of insufficient evidence.

Table 8. Numbers of exons and transcripts supported by various types of evidence for Otto and de novo gene prediction methods. Highlighted cells indicate the gene sets analyzed in this paper (boldface, set of genes selected for protein analysis; italic, total set of accepted de novo predictions).

		Total	Types of evidence				No. of lines of evidence*			
			Mouse	Rodent	Protein	Human	≥1	≥2	≥3	≥4
Otto	Number of transcripts	17,969	17,065	14,881	15,477	16,374	**17,968†**	17,501	15,877	12,451
	Number of exons	141,218	111,174	89,569	108,431	118,869	140,710	127,955	99,574	59,804
De novo	Number of transcripts	58,032	14,463	5,094	8,043	9,220	*21,350*	**8,619**	4,947	1,904
	Number of exons	319,935	48,594	19,344	26,264	40,104	79,148	31,130	17,508	6,520
No. of exons per transcript	Otto	7.84	5.77	6.01	6.99	7.24	7.81	7.19	6.00	4.28
	De novo	5.53	3.17	3.80	3.27	4.36	3.7	3.56	3.42	3.16

*Four kinds of evidence (conservation in 3× mouse genomic DNA, similarity to human EST or cDNA, similarity to rodent EST or cDNA, and similarity to known proteins) were considered to support gene predictions from the different methods. The use of evidence is quite liberal, requiring only a partial match to a single exon of predicted transcript. †This number includes alternative splice forms of the 17,764 genes mentioned elsewhere in the text.

Examination of pericentromeric regions is ongoing.

The remaining ~80% of the genome, the euchromatic component, is divisible into G-, R-, and T-bands (*67*). These cytogenetic bands have been presumed to differ in their nucleotide composition and gene density, although we have been unable to determine precise band boundaries at the molecular level. T-bands are the most G+C- and gene-rich, and G-bands are G+C-poor (*68*). Bernardi has also offered a description of the euchromatin at the molecular level as long stretches of DNA of differing base composition, termed isochores (denoted L, H1, H2, and H3), which are >300 kbp in length (*69*). Bernardi defined the L (light) isochores as G+C-poor (<43%), whereas the H (heavy) isochores fall into three G+C-rich classes representing 24, 8, and 5% of the genome. Gene concentration has been claimed to be very low in the L isochores and 20-fold more enriched in the H2 and H3 isochores (*70*). By examining contiguous 50-kbp windows of G+C content across the assembly, we found that regions of G+C content >48% (H3 isochores) averaged 273.9 kbp in length, those with G+C content between 43 and 48% (H1+H2 isochores) averaged 202.8 kbp in length, and the average span of regions with <43% (L isochores) was 1078.6 kbp. The correlation between G+C content and gene density was also examined in 50-kbp windows along the assembled sequence (Table 9 and Figs. 10 and 11). We found that the density of genes was greater in regions of high G+C than in regions of low G+C content, as expected. However, the correlation between G+C content and gene density was not as skewed as previously predicted (*69*). A higher proportion of genes were located in the G+C-poor regions than had been expected.

Chromosomes 17, 19, and 22, which have a disproportionate number of H3-containing bands, had the highest gene density (Table 10). Conversely, of the chromosomes that we found to have the lowest gene density, X, 4, 18, 13, and Y, also have the fewest H3 bands. Chromosome 15, which also has few H3 bands, did not have a particularly low gene density in our analysis. In addition, chromosome 8, which we found to have a low gene density, does not appear to be unusual in its H3 banding.

How valid is Ohno's postulate (*71*) that mammalian genomes consist of oases of genes in otherwise essentially empty deserts? It appears that the human genome does indeed contain deserts, or large, gene-poor regions. If we define a desert as a region >500 kbp without a gene, then we see that 605 Mbp, or about 20% of the genome, is in deserts. These are not uniformly distributed over the various chromosomes. Gene-rich chromosomes 17, 19, and 22 have only about 12% of their collective 171 Mbp in deserts, whereas gene-poor chromosomes 4, 13, 18, and X have 27.5% of their 492 Mbp in deserts (Table 11). The apparent lack of predicted genes in these regions does not necessarily imply that they are devoid of biological function.

4.2 Linkage map

Linkage maps provide the basis for genetic analysis and are widely used in the study of the inheritance of traits and in the positional cloning of genes. The distance metric, centimorgans (cM), is based on the recombination rate between homologous chromosomes during meiosis. In general, the rate of recombination in females is greater than that in males, and this degree of map expansion is not uniform across the genome (*72*). One of the opportunities enabled by a nearly complete genome sequence is to produce the ultimate physical map, and to fully analyze its correspondence with two other maps that have been widely used in genome and genetic analysis: the linkage map and the cytogenetic map. This would close the loop between the mapping and sequencing phases of the genome project.

We mapped the location of the markers that constitute the Genethon linkage map to the genome. The rate of recombination, expressed as cM per Mbp, was calculated for 3-Mbp windows as shown in Table 12. Higher rates of recombination in the telomeric region of the chromosomes have been previously documented (*73*). From this mapping result, there is a difference of 4.99 between lowest rates and highest rates and the largest difference of 4.4 between males and females (4.99 to 0.47 on chromosome 16). This indicates that the variability in recombination rates among regions of the genome exceeds the differences in recombination rates between males and females. The human genome has recombination hotspots, where recombination rates vary fivefold or more over a space of 1 kbp, so the picture one gets of the magnitude of variability in recombination rate will depend on the size of the window

Table 9. Characteristics of G+C in isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43–48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

*The predictions were based on Bernardi's definitions (*70*) of the isochore structure of the human genome.

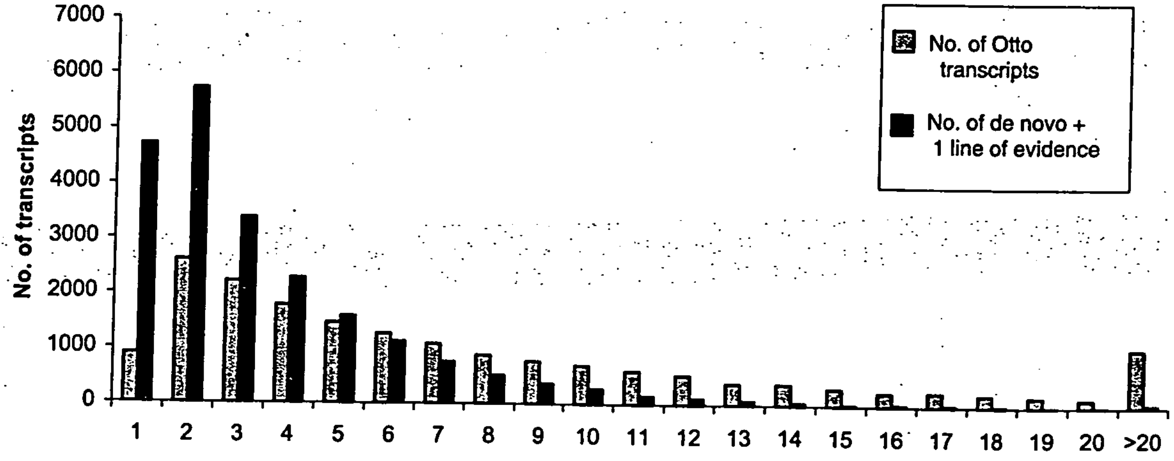


Fig. 9. Comparison of the number of exons per transcript between the 17,968 Otto transcripts and 21,350 de novo transcript predictions with at least one line of evidence that do not overlap with an Otto prediction. Both sets have the highest number of transcripts in the two-exon category, but the de novo gene predictions are skewed much more toward smaller transcripts. In the Otto set, 19.7% of the transcripts have one or two exons, and 5.7% have more than 20. In the de novo set, 49.3% of the transcripts have one or two exons, and 0.2% have more than 20.

examined. Unfortunately, too few meiotic crossovers have occurred in Centre d'Étude du Polymorphism Humain (CEPH) and other reference families to provide a resolution any finer than about 3 Mbp. The next challenge will be to determine a sequence basis of recombination at the chromosomal level. An accurate predictor for the rate for variation in recombination rates between any pair of markers would be extremely useful in designing markers to narrow a region of linkage, such as in positional cloning projects.

4.3 Correlation between CpG islands and genes

CpG islands are stretches of unmethylated DNA with a higher frequency of CpG dinucleotides when compared with the entire genome (*74*). CpG islands are believed to preferentially occur at the transcriptional start of genes, and it has been observed that most housekeeping genes have CpG islands at the 5′ end of the transcript (*75, 76*). In addition, experimental evidence indicates that CpG island methylation is correlated with gene inactivation (*77*) and has been shown to be important during gene imprinting (*78*) and tissue-specific gene expression (*79*)

Experimental methods have been used that resulted in an estimate of 30,000 to 45,000 CpG islands in the human genome (*74, 80*) and an estimate of 499 CpG islands on human chromosome 22 (*81*). Larsen *et al.* (*76*) and Gardiner-Garden and Frommer (*75*) used a computational method to identify CpG islands and defined them as regions of DNA of >200 bp that have a G+C content of >50% and a ratio of observed versus expected frequency of CG dinucleotide ≥0.6.

It is difficult to make a direct comparison of experimental definitions of CpG islands with computational definitions because computational methods do not consider the methylation state of cytosine and experimental methods do not directly select regions of high G+C content. However, we can determine the correlation of CpG island with gene starts, given a set of annotated genomic transcripts and the whole genome sequence. We have analyzed the publicly available annotation of chromosome 22, as well as using the entire human genome in our assembly and the computationally annotated genes. A variation of the CpG island computation was compared with Larsen *et al.* (*76*). The main differences are that we use a sliding window of 200 bp, consecutive windows are merged only if they overlap, and we recompute the CpG value upon merging, thus rejecting any potential island if it scores less than the threshold.

To compute various CpG statistics, we used two different thresholds of CG dinucleotide likelihood ratio. Besides using the original threshold of 0.6 (method 1), we used a higher threshold of CG dinucleotide likelihood ratio of 0.8 (method 2), which results in the number of CpG islands on chromosome 22 close to the number of annotated genes on this chromosome. The main results are summarized in Table 13. CpG islands computed with method 1 predicted only 2.6% of the CSA sequence as CpG, but 40% of the gene starts (start codons) are contained inside a CpG island. This is comparable to ratios reported by others (*82*). The last two rows of the table show the observed and expected average distance, respectively, of the closest CpG island from the first exon. The observed average closest CpG islands are smaller than the corresponding expected distances, confirming an association between CpG island and the first exon.

We also looked at the distribution of CpG island nucleotides among various sequence classes such as intergenic regions, introns, exons, and first exons. We computed the likelihood score for each sequence class as the ratio of the observed fraction of CpG island nucleotides in that sequence class and the expected fraction of CpG island nucleotides in that sequence class. The result of applying method 1 on CSA were scores of 0.89 for intergenic region, 1.2 for intron, 5.86 for exon, and 13.2 for first exon. The same trend was also found for chromosome 22 and after the application of a higher threshold (method 2) on both data sets. In sum, genome-wide analysis has extended earlier analysis and suggests a strong correlation between CpG islands and first coding exons.

4.4 Genome-wide repetitive elements

The proportion of the genome covered by various classes of repetitive DNA is presented in Table 14. We observed about 35% of the genome in these repeat classes, very similar to values reported previously (*83*). Repetitive sequence may be underrepresented in the Celera assembly as a result of incomplete repeat resolution, as discussed above. About 8% of the scaffold length is in gaps, and we expect that much of this is repetitive sequence. Chromosome 19 has the highest repeat density (57%), as well as the highest gene density (Table 10). Of interest, among the different classes of repeat elements, we observe a clear association of Alu elements and gene density, which was not observed between LINEs and gene density.

5 Genome Evolution

Summary. The dynamic nature of genome evolution can be captured at several levels. These include gene duplications mediated by RNA intermediates (retrotransposition) and segmental genomic duplications. In this section, we document the genome-wide occurrence of retrotransposition events generating functional (intronless paralogs) or inactive genes (pseudogenes). Genes involved in translational processes and nuclear regulation account for nearly 50% of all intronless paralogs and processed pseudogenes detected in our survey. We have also cataloged the extent of segmental genomic duplication and provide evidence for 1077 duplicated blocks covering 3522 distinct genes.

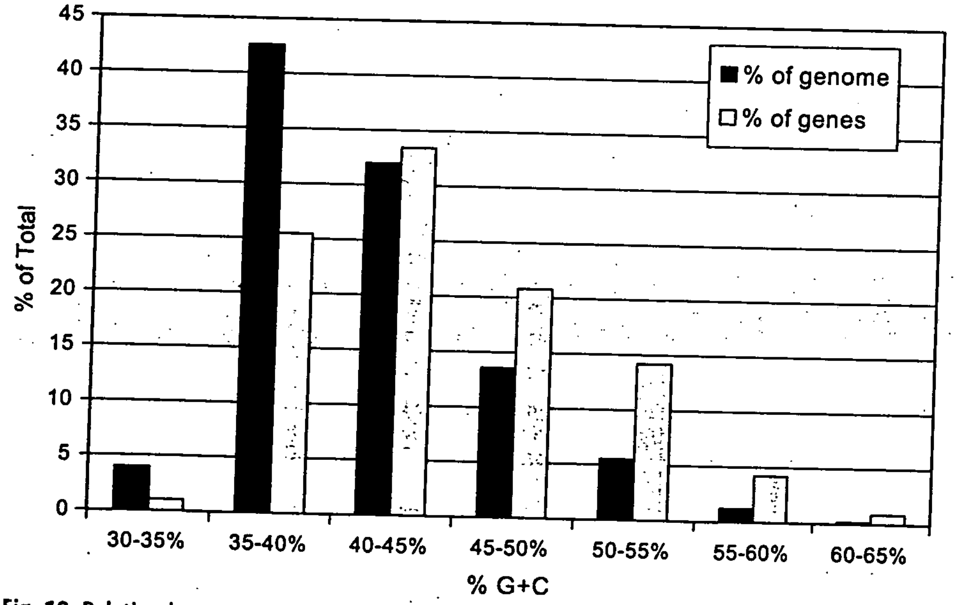


Fig. 10. Relation between G+C content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated G+C content. The percent of the total number of genes associated with each G+C bin is represented by the yellow bars. The graph shows that about 5% of the genome has a G+C content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

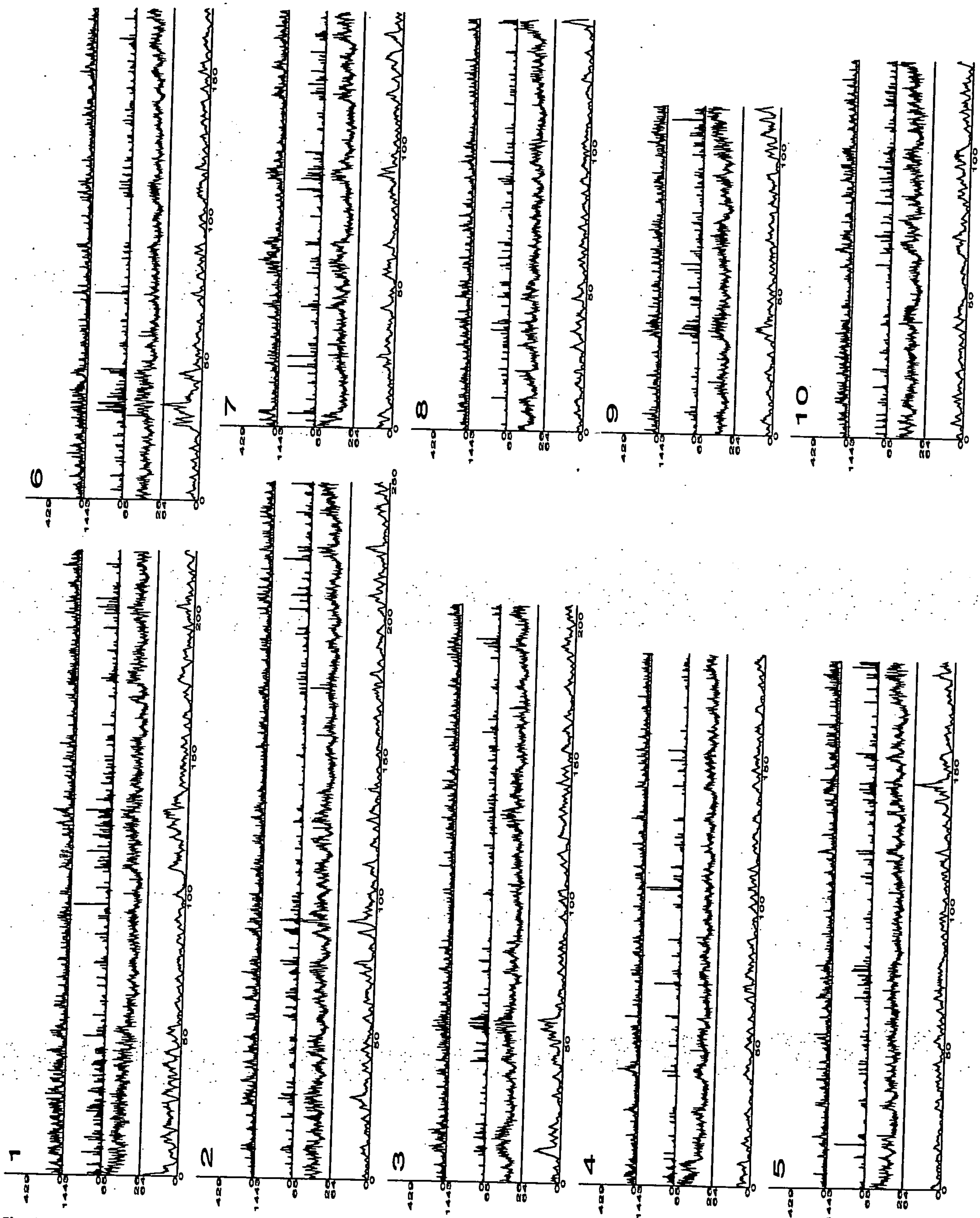


Fig. 11. Genome structural features.

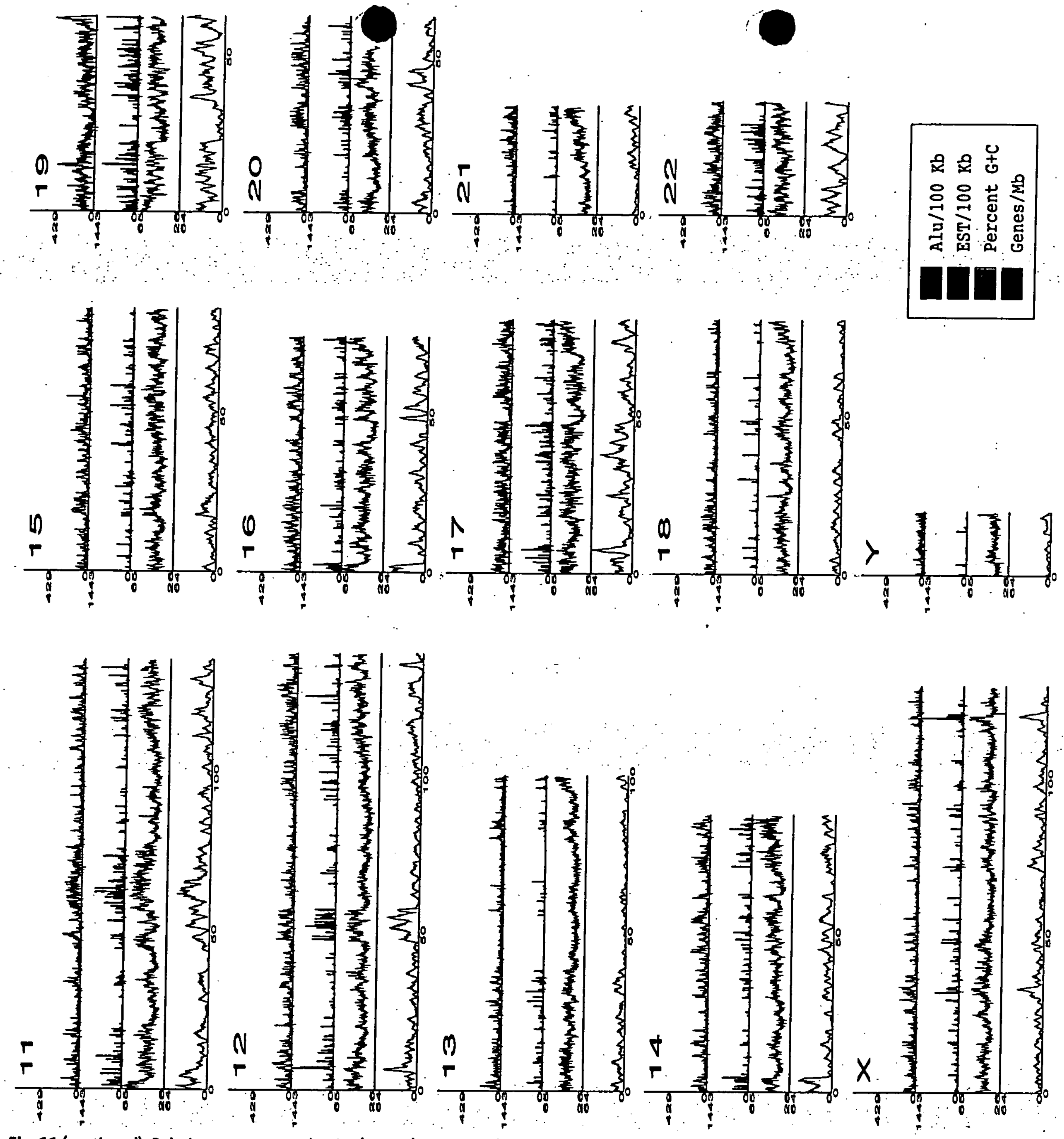


Fig. 11 (continued). Relation among gene density (orange), G+C content (green), EST density (blue), and Alu density (pink) along the lengths of each of the chromosomes. Gene density was calculated in 1-Mbp windows. The percent of G+C nucleotides was calculated in 100-kbp windows. The number of ESTs and Alu elements is shown per 100-kbp window.

5.1 Retrotransposition in the human genome

Retrotransposition of processed mRNA transcripts into the genome results in functional genes, called intronless paralogs, or inactivated genes (pseudogenes). A paralog refers to a gene that appears in more than one copy in a given organism as a result of a duplication event. The existence of both intron-containing and intronless forms of genes encoding functionally similar or identical proteins has been previously described (*84*, *85*). Cataloging these evolutionary events on the genomic landscape is of value in understanding the functional consequences of such gene-duplication events in cellular biology. Identification of conserved intronless paralogs in the mouse or other mammalian genomes should provide the basis for capturing the evolutionary chronology of these transposition events and provide insights into gene loss and accretion in the mammalian radiation.

A set of proteins corresponding to all 901

Table 10. Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2x refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaffolds	Largest scaffold (Mbp)	No. of scaffolds >500 kbp	Sequence covered by scaffolds >500 kbp	% of total sequence in scaffolds >500 kbp	% repeat	% GC	No of CpG islands	Otto	De novo/ any	De novo/ 2×	Total (Otto + de novo/ any)	Total (Otto + de novo/ any)	Sequence in deserts >500/ kbp	Sequence in deserts >1 Mbp	Otto	De novo/ any	De novo/ 2×	Otto + de novo/ any	Otto + de novo/ 2×
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

*Chromosomal assignment unknown.

Otto-predicted, single-exon genes were subjected to BLAST analysis against the proteins encoded by the remaining multiexon predicted transcripts. Using homology criteria of 70% sequence identity over 90% of the length, we identified 298 instances of single- to multi-exon correspondence. Of these 298 sequences, 97 were represented in the GenBank data set of experimentally validated full-length genes at the stringency specified and were verified by manual inspection.

We believe that these 97 cases may represent intronless paralogs (see Web table 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) of known genes. Most of these are flanked by direct repeat sequences, although the precise nature of these repeats remains to be determined. All of the cases for which we have high confidence contain polyadenylated [poly(A)] tails characteristic of retrotransposition.

Recent publications describing the phenomenon of functional intronless paralogs speculate that retrotransposition may serve as a mechanism used to escape X-chromosomal inactivation (*84*, *86*). We do not find a bias toward X chromosome origination of these retrotransposed genes; rather, the results show a random chromosome distribution of both the intron-containing and corresponding intronless paralogs. We also have found several cases of retrotransposition from a single source chromosome to multiple target chromosomes. Interesting examples include the retrotransposition of a five exon–containing ribosomal protein L21 gene on chromosome 13 onto chromosomes 1, 3, 4, 7, 10, and 14, respectively. The size of the source genes can also show variability. The largest example is the 31-exon diacylglycerol kinase zeta gene on chromosome 11 that has an intronless paralog on chromosome 13. Regardless of route, retrotransposition with subsequent gene changes in coding or noncoding regions that lead to different functions or expression patterns, represents a key route to providing an enhanced functional repertoire in mammals (*87*).

Our preliminary set of retrotransposed intronless paralogs contains a clear overrepresentation of genes involved in translational processes (40% ribosomal proteins and 10% translation elongation factors) and nuclear regulation (HMG nonhistone proteins, 4%), as well as metabolic and regulatory enzymes. EST matches specific to a subset of intronless paralogs suggest expression of these intronless paralogs. Differences in the upstream regulatory sequences between the source genes and their intronless paralogs could account for differences in tissue-specific gene expression. Defining which, if any, of these processed genes are functionally expressed and translated will require further elucidation and experimental validation.

5.2 Pseudogenes

A pseudogene is a nonfunctional copy that is very similar to a normal gene but that has been altered slightly so that it is not expressed. We developed a method for the preliminary analysis of processed pseudogenes in the human genome as a starting point in elucidating the ongoing evolutionary forces

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

*In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Table 12. Rate of recombination per physical distance (cM/Mb) across the genome. Genethon markers were placed on CSA-mapped assemblies, and then relative physical distances and rates were calculated in 3-Mb windows for each chromosome. NA, not applicable.

Chrom.	Male			Sex-average			Female		
	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.
1	2.60	1.12	0.23	2.81	1.42	0.52	3.39	1.76	0.68
2	2.23	0.78	0.33	2.65	1.12	0.54	3.17	1.40	0.61
3	2.55	0.86	0.23	2.40	1.07	0.42	2.71	1.30	0.33
4	1.66	0.67	0.15	2.06	1.04	0.60	2.50	1.40	0.77
5	2.00	0.67	0.18	1.87	1.08	0.42	2.26	1.43	0.62
6	1.97	0.71	0.28	2.57	1.12	0.37	3.47	1.67	0.64
7	2.34	1.16	0.48	1.67	1.17	0.47	2.27	1.21	0.34
8	1.83	0.73	0.14	2.40	1.05	0.46	3.44	1.36	0.43
9	2.01	0.99	0.53	1.95	1.32	0.77	2.63	1.66	0.82
10	3.73	1.03	0.22	3.05	1.29	0.66	2.84	1.51	0.76
11	1.43	0.72	0.31	2.13	0.99	0.47	3.10	1.32	0.49
12	4.12	0.76	0.26	3.35	1.16	0.49	2.93	1.55	0.59
13	1.60	0.75	0.01	1.87	0.95	0.17	2.49	1.19	0.32
14	3.15	0.98	0.18	2.65	1.30	0.62	3.14	1.63	0.75
15	2.28	0.94	0.34	2.31	1.22	0.42	2.53	1.56	0.54
16	1.83	1.00	0.47	2.70	1.55	0.63	4.99	2.32	1.12
17	3.87	0.87	0.00	3.54	1.35	0.54	4.19	1.83	0.94
18	3.12	1.37	0.86	3.75	1.66	0.43	4.35	2.24	0.72
19	3.02	0.97	0.10	2.57	1.41	0.49	2.89	1.75	0.87
20	3.64	0.89	0.00	2.79	1.50	0.83	3.31	2.15	1.34
21	3.23	1.26	0.69	2.37	1.62	1.08	2.58	1.90	1.18
22	1.25	1.10	0.84	1.88	1.41	1.08	3.73	2.08	0.93
X	NA	NA	NA	NA	NA	NA	3.12	1.64	0.72
Y	NA	NA	NA	NA	NA	NA	NA	NA	NA
Genome	4.12	0.88	0.00	3.75	1.22	0.17	4.99	1.55	0.32

that account for gene inactivation. The general structural characteristics of these processed pseudogenes include the complete lack of intervening sequences found in the functional counterparts, a poly(A) tract at the 3′ end, and direct repeats flanking the pseudogene sequence. Processed pseudogenes occur as a result of retrotransposition, whereas unprocessed pseudogenes arise from segmental genome duplication.

We searched the complete set of Otto-predicted transcripts against the genomic sequence by means of BLAST. Genomic regions corresponding to all Otto-predicted transcripts were excluded from this analysis. We identified 2909 regions matching with greater than 70% identity over at least 70% of the length of the transcripts that likely represent processed pseudogenes. This number is probably an underestimate because specific methods to search for pseudogenes were not used.

We looked for correlations between structural elements and the propensity for retrotransposition in the human genome. GC content and transcript length were compared between the genes with processed pseudogenes (1177 source genes) versus the remainder of the predicted gene set. Transcripts that give rise to processed pseudogenes have shorter average transcript length (1027 bp versus 1594 bp for the Otto set) as compared with genes for which no pseudogene was detected. The overall GC content did not show any significant difference, contrary to a recent report (*88*). There is a clear trend in gene families that are present as processed pseudogenes. These include ribosomal proteins (67%), lamin receptors (10%), translation elongation factor alpha (5%), and HMG–non-histone proteins (2%). The increased occurrence of retrotransposition (both intronless paralogs and processed pseudogenes) among genes involved in translation and nuclear regulation may reflect an increased transcriptional activity of these genes.

5.3 Gene duplication in the human genome

Building on a previously published procedure (*27*), we developed a graph-theoretic algorithm, called Lek, for grouping the predicted human protein set into protein families (*89*).

Table 13. Characteristics of CpG islands identified in chromosome 22 (34-Mbp sequence length) and the whole genome (2.9-Gbp sequence length) by means of two different methods. Method 1 uses a CG likelihood ratio of ≥0.6. Method 2 uses a CG likelihood ratio of ≥0.8.

	Chromosome 22		Whole genome (CS assembly)	
	Method 1	Method 2	Method 1	Method 2
Number of CpG islands detected	5,211	522	195,706	26,876
Average length of island (bp)	390	535	395	497
Percent of sequence predicted as CpG	5.9	0.8	2.6	0.4
Percent of first exons that overlap a CpG island	44	25	42	22
Percent of first exons with first position of exon contained inside a CpG island	37	22	40	21
Average distance between first exon and closest CpG island (bp)	1,013	10,486	2,182	17,021
Expected distance between first exon and closest CpG island (bp)	3,262	32,567	7,164	55,811

Table 14. Distribution of repetitive DNA in the compartmentalized shotgun assembly sequence.

Repetitive elements	Megabases in assembled sequences	Percent of assembly	Previously predicted (%) (*83*)
Alu	288	9.9	10.0
Mammalian interspersed repeat (MIR)	66	2.3	1.7
Medium reiteration (MER)	50	1.7	1.6
Long terminal repeat (LTR)	155	5.3	5.6
Long interspersed nucleotide element (LINE)	466	16.1	16.7
Total	1025	35.3	35.6

The complete clusters that result from the Lek clustering provide one basis for comparing the role of whole-genome or chromosomal duplication in protein family expansion as opposed to other means, such as tandem duplication. Because each complete cluster represents a closed and certain island of homology, and because Lek is capable of simultaneously clustering protein complements of several organisms, the number of proteins contributed by each organism to a complete cluster can be predicted with confidence depending on the quality of the annotation of each genome. The variance of each organism's contribution to each cluster can then be calculated, allowing an assessment of the relative importance of large-scale duplication versus smaller-scale, organism-specific expansion and contraction of protein families, presumably as a result of natural selection operating on individual protein families within an organism. As can be seen in Fig. 12, the large variance in the relative numbers of human as compared with *D. melanogaster* and *Caenorhabditis elegans* proteins in complete clusters may be explained by multiple events of relative expansions in gene families in each of the three animal genomes. Such expansions would give rise to the distribution that shows a peak at 1:1 in the ratio for human-worm or human-fly clusters with the slope spread covering both human and fly/worm predominance, as we observed (Fig. 12). Furthermore, there are nearly as many clusters where worm and fly proteins predominate despite the larger numbers of proteins in the human. At face value, this analysis suggests that natural selection acting on individual protein families has been a major force driving the expansion of at least some elements of the human protein set. However, in our analysis, the difference between an ancient whole-genome duplication followed by loss, versus piecemeal duplication, cannot be easily distinguished. In order to differentiate these scenarios, more extended analyses were performed.

5.4 Large-scale duplications

Using two independent methods, we searched for large-scale duplications in the human genome. First, we describe a protein family–based method that identified highly conserved blocks of duplication. We then describe our comprehensive method for identifying all interchromosomal block duplications. The latter method identified a large number of duplicated chromosomal segments covering parts of all 24 chromosomes.

The first of the methods is based on the idea of searching for blocks of highly conserved homologous proteins that occur in more than one location on the genome. For this comparison, two genes were considered equivalent if their protein products were determined to be in the same family and the same complete Lek cluster (essentially paralogous genes) (89). Initially, each chromosome was represented as a string of genes ordered by the start codons for predicted genes along the chromosome. We considered the two strands as a single string, because local inversions are relatively common events relative to large-scale duplications. Each gene was indexed according to the protein family and Lek complete cluster (89). All pairs of indexed gene strings were then aligned in both the forward and reverse directions with the Smith-Waterman algorithm (90). A match between two proteins of the same Lek complete cluster was given a score of 10 and a mismatch −10, with gap open and extend penalties of −4 and −1. With these parameters, 19 conserved interchromosomal blocks of duplication were observed, all of which were also detected and expanded by the comprehensive method described below. The detection of only a relatively small number of block duplications was a consequence of using an intrinsically conservative method grounded in the conservative constraints of the complete Lek clusters.

In the second, more comprehensive approach, we aligned all chromosomes directly with one another using an algorithm based on the MUMmer system (91). This alignment method uses a suffix tree data structure and a linear-time algorithm to align long sequences very rapidly; for example, two chromosomes of 100 Mbp can be aligned in less than 20 min (on a Compaq Alpha computer) with 4 gigabytes of memory. This procedure was used recently to identify numerous large-scale segmental duplications among the five chromosomes of *A. thaliana* (92); in that organism, the method revealed that 60% of the genome (66 Mbp) is covered by 24 very large duplicated segments. For *Arabidopsis*, a DNA-based alignment was sufficient to reveal the segmental duplications between chromosomes; in the human genome, DNA alignments at the whole-chromosome level are insufficiently sensitive. Therefore, a modified procedure was developed and applied, as follows. First, all 26,588 proteins (9,675,713 million amino acids) were concatenated end-to-end in order as they occur along each of the 24 chromosomes, irrespective of strand location. The concatenated protein set was then aligned against each chromosome by the MUMmer algorithm. The resulting matches were clustered to extract all sets of three or more protein matches that occur in close proximity on two different chromosomes (93); these represent the candidate segmental duplications. A series of filters were developed and applied to remove likely false-positives from this set; for example, small blocks that were spread across many proteins were removed. To refine the filtering methods, a shuffled protein set was first created by taking the 26,588 proteins, randomizing their order, and then partitioning them into 24 shuffled chromosomes, each containing the same number of proteins as the true genome. This shuffled protein set has the identical composition to the real genome; in particular, every protein and every domain appears the same number of times. The complete algorithm was then applied to both the real and the shuffled data, with the results on the shuffled data being used to estimate the false-positive rate. The algorithm after filtering yielded 10,310 gene pairs in 1077 duplicated blocks containing 3522 distinct genes; tandemly duplicated expansions in many of the blocks explain the excess of gene pairs to distinct genes. In the shuffled data, by contrast, only 370 gene pairs were found, giving a false-positive estimate of 3.6%. The most likely explanation for the 1077 block duplications is ancient segmental duplications. In many cases, the order of the proteins has been shuffled, although proximity is preserved. Out of the 1077 blocks, 159 contain only three genes, 137 contain four genes, and 781 contain five or more genes.

To illustrate the extent of the detected duplications, Fig. 13 shows all 1077 block duplications indexed to each chromosome in 24 panels in which only duplications mapped to the indexed chromosome are displayed. The figure makes it clear that the duplications are ubiquitous in the genome. One feature that it displays is many relatively small chromosomal stretches, with one-to-many duplication relationships that are graphically striking. One such example captured by the analysis is the well-documented olfactory receptor (OR) family, which is scattered in blocks throughout the genome and which has been analyzed for genome-deployment reconstruction at several evolutionary stages (94). The figure also illustrates that some chromosomes, such as chromosome 2, contain many more detected large-scale duplications than others. Indeed, one of the largest duplicated segments is a large block of 33 proteins on chromosome 2, spread among eight smaller blocks in 2p, that aligns to a paralogous set on chromosome 14, with one rearrangement (see chromosomes 2 and 14 panels in Fig. 13). The proteins are not contiguous but span a region containing 97 proteins on chromosome 2 and 332 proteins on chromosome 14. The likelihood of observing this many duplicated proteins by chance, even over a span of this length, is 2.3×10^{-68} (93). This duplicated set spans 20 Mbp on chromosome 2 and 63 Mbp on chromosome 14, over 70% of the latter chromosome. Chromosome 2 also contains a block duplication that is nearly as large, which is shared by chromosome arm 2q and chromosome 12. This duplication incorporates two of the four known Hox gene clusters, but considerably expands the extent of the duplications proximally and distally on the pair of chromosome arms. This breadth of duplication is also seen on the two chromosomes carrying the other two Hox clusters.

An additional large duplication, between chromosomes 18 and 20, serves as a good example to illustrate some of the features common to many of the other observed large duplications (Fig. 13, inset). This duplication contains 64 detected ordered intrachromosomal pairs of homologous genes. After discounting a 40-Mb stretch of chromosome 18 free of matches to chromosome 20, which is likely to represent a large insert (between the gene assignments "Krup rel" and "collagen rel" on chromosome 18 in Fig. 13), the full duplication segment covers 36 Mb on chromosome 18 and 28 Mb on chromosome 20.

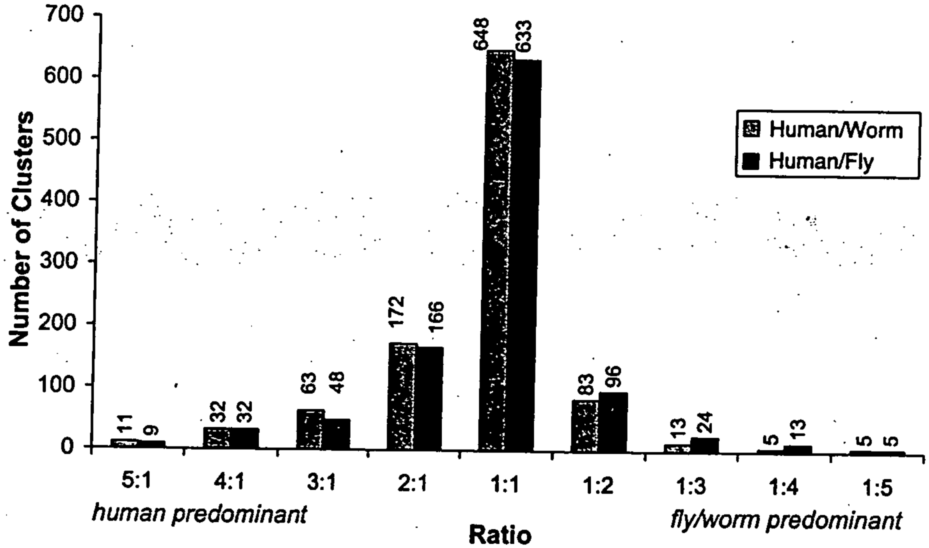


Fig. 12. Gene duplication in complete protein clusters. The predicted protein sets of human, worm, and fly were subjected to Lek clustering (27). The numbers of clusters with varying ratios (whole number) of human versus worm and human versus fly proteins per cluster were plotted.

By this measure, the duplication segment spans nearly half of each chromosome's net length. The most likely scenario is that the whole span of this region was duplicated as a single very large block, followed by shuffling owing to smaller scale rearrangements. As such, at least four subsequent rearrangements would need to be invoked to explain the relative insertions and inversions seen in the duplicated segment interval. The 64 protein pairs in this alignment occur among 217 protein assignments on chromosome 18, and among 322 protein assignments on chromosome 20, for a density of involved proteins of 20 to 30%. This is consistent with an ancient large-scale duplication followed by subsequent gene loss on one or both chromosomes. Loss of just one member of a gene pair subsequent to the duplication would result in a failure to score a gene pair in the block; less than 50% gene loss on the chromosomes would lead to the duplication density observed here. As an independent verification of the significance of the alignments detected, it can be seen that a substantial number of the pairs of aligning proteins in this duplication, including some of those annotated (Fig. 13), are those populating small Lek complete clusters (see above). This indicates that they are members of very small families of paralogs; their relative scarcity within the genome validates the uniqueness and robust nature of their alignments.

Two additional qualitative features were observed among many of the large-scale duplications. First, several proteins with disease associations, with OMIM (Online Mendelian Inheritance in Man) assignments, are members of duplicated segments (see web table 2 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1). We have also observed a few instances where paralogs on both duplicated segments are associated with similar disease conditions. Notable among these genes are proteins involved in hemostasis (coagulation factors) that are associated with bleeding disorders, transcriptional regulators like the homeobox proteins associated with developmental disorders, and potassium channels associated with cardiovascular conduction abnormalities. For each of these disease genes, closer study of the paralogous genes in the duplicated segment may reveal new insights into disease causation, with further investigation needed to determine whether they might be involved in the same or similar genetic diseases. Second, although there is a conserved number of proteins and coding exons predicted for specific large duplicated spans within the chromosome 18 to 20 alignment, the genomic DNA of chromosome 18 in these specific spans is in some cases more than 10-fold longer than the corresponding chromosome 20 DNA. This selective accretion of noncoding DNA (or conversely, loss of noncoding DNA) on one of a pair of duplicated chromosome regions was observed in many compared regions. Hypotheses to explain which mechanisms foster these processes must be tested.

Evaluation of the alignment results gives some perspective on dating of the duplications. As noted above, large-scale ancient segmental duplication in fact best explains many of the blocks detected by this genome-wide analysis. The regions of human chromosomes involved in the large-scale duplications expanded upon above (chromosomes 2 to 14, 2 to 12, and 18 to 20) are each syntenic to a distinct mouse chromosomal region. The corresponding mouse chromosomal regions are much more similar in sequence conservation, and even in order, to their human synteny partners than the human duplication regions are to each other. Further, the corresponding mouse chromosomal regions each bear a significant proportion of genes orthologous to the human genes on which the human duplication assignments were made. On the basis of these factors, the corresponding mouse chromosomal spans, at coarse resolution, appear to be products of the same large-scale duplications observed in humans. Although further detailed analysis must be carried out once a more complete genome is assembled for mouse, the underlying large duplications appear to predate the two species' divergence. This dates the duplications, at the latest, before divergence of the primate and rodent lineages. This date can be further refined upon examination of the synteny between human chromosomes and those of chicken, pufferfish (*Fugu rubripes*), or zebrafish (*95*). The only substantial syntenic stretches mapped in these species corresponding to both pairs of human duplications are restricted to the Hox cluster regions. When the synteny of these regions (or others) to human chromosomes is extended with further mapping, the ages of the nearly chromosome-length duplications seen in humans are likely to be dated to the root of vertebrate divergence.

The MUMmer-based results demonstrate large block duplications that range in size from a few genes to segments covering most of a chromosome. The extent of segmental duplications raises the question of whether an ancient whole-genome duplication event is the underlying explanation for the numerous duplicated regions (*96*). The duplications have undergone many deletions and subsequent rearrangements; these events make it difficult to distinguish between a whole-genome duplication and multiple smaller events. Further analysis, focused especially on comparing the estimated ages of all the block duplications, derived partially from interspecies genome comparisons, will be necessary to determine which of these two hypotheses is more likely. Comparisons of genomes of different vertebrates, and even cross-phyla genome comparisons, will allow for the deconvolution of duplications to eventually reveal the stagewise history of our genome, and with it a history of the emergence of many of the key functions that distinguish us from other living things.

6 A Genome-Wide Examination of Sequence Variations

Summary. Computational methods were used to identify single-nucleotide polymorphisms (SNPs) by comparison of the Celera sequence to other SNP resources. The SNP rate between two chromosomes was ~1 per 1200 to 1500 bp. SNPs are distributed nonrandomly throughout the genome. Only a very small proportion of all SNPs (<1%) potentially impact protein function based on the functional analysis of SNPs that affect the predicted coding regions. This results in an estimate that only thousands, not millions, of genetic variations may contribute to the structural diversity of human proteins.

Having a complete genome sequence enables researchers to achieve a dramatic acceleration in the rate of gene discovery, but only through analysis of sequence variation in DNA can we discover the genetic basis for variation in health among human beings. Whole-genome shotgun sequencing is a particularly effective method for detecting sequence variation in tandem with whole-genome assembly. In addition, we compared the distribution and attributes of SNPs ascertained by three other methods: (i) alignment of the Celera consensus sequence to the PFP assembly, (ii) overlap of high-quality reads of genomic sequence (referred to as "Kwok"; 1,120,195 SNPs) (*97*), and (iii) reduced representation shotgun sequencing (referred to as "TSC"; 632,640 SNPs) (*98*). These data were consistent in showing an overall nucleotide diversity of $\sim 8 \times 10^{-4}$, marked heterogeneity across the genome in SNP density, and an overwhelming preponderance of noncoding variation that produces no change in expressed proteins.

6.1 SNPs found by aligning the Celera consensus to the PFP assembly

Ideally, methods of SNP discovery make full use of sequence depth and quality at every site, and quantitatively control the rate of false-positive and false-negative calls with an explicit sampling model (*99*). Comparison of consensus sequences in the absence of these details necessitated a more ad hoc approach (quality scores could not readily be obtained for the PFP assembly). First, all sequence differences between the two consensus sequences were identified; these were then filtered to reduce the contribution of sequencing errors and misassembly. As a measure of the effectiveness of the filtering step, we monitored the ratio of transition and transversion substitutions, because a 2:1 ratio has been well documented as typical in mammalian evolution (*100*) and in human SNPs

(*101, 102*). The filtering steps consisted of removing variants where the quality score in the Celera consensus was less than 30 and where the density of variants was greater than 5 in 400 bp. These filters resulted in shifting the transition-to-transversion ratio from 1.57:1 to 1.89:1. When applied to 2.3 Gbp of alignments between the Celera and PFP consensus sequences, these filters resulted in identification of 2,104,820 putative SNPs from a total of 2,778,474 substitution differences. Overlaps between this set of SNPs and those found by other methods are described below.

6.2 Comparisons to public SNP databases

Additional SNPs, including 2,536,021 from dbSNP (www.ncbi.nlm.nih.gov/SNP) and 13,150 from HGMD (Human Gene Mutation Database, from the University of Wales, UK), were mapped on the Celera consensus sequence by a sequence similarity search with the program PowerBlast (*103*). The two largest data sets in dbSNP are the Kwok and TSC sets, with 47% and 25% of the dbSNP records. Low-quality alignments with partial coverage of the dbSNP sequence and alignments that had less than 98% sequence identity between the Celera sequence and the dbSNP flanking sequence were eliminated. dbSNP sequences mapping to multiple locations on the Celera genome were discarded. A total of 2,336,935 dbSNP variants were mapped to 1,223,038 unique locations on the Celera sequence, implying considerable redundancy in dbSNP. SNPs in the TSC set mapped to 585,811 unique genomic locations, and SNPs in the Kwok set mapped to 438,032 unique locations. The combined unique SNPs counts used in this analysis, including Celera-PFP, TSC, and Kwok, is 2,737,668. Table 15 shows that a substantial fraction of SNPs identified by one of these methods was also found by another method. The very high overlap (36.2%) between the Kwok and Celera-PFP SNPs may be due in part to the use by Kwok of sequences that went into the PFP assembly. The unusually low overlap (16.4%) between the Kwok and TSC sets is due to their being the smallest two sets. In addition, 24.5% of the Celera-PFP SNPs overlap with SNPs derived from the Celera genome sequences (*46*). SNP validation in population samples is an expensive and laborious process, so confirmation on multiple data sets may provide an efficient initial validation "in silico" (by computational analysis).

Table 15. Overlap of SNPs from genome-wide SNP databases. Table entries are SNP counts for each pair of data sets. Numbers in parentheses are the fraction of overlap, calculated as the count of overlapping SNPs divided by the number of SNPs in the smaller of the two databases compared. Total SNP counts for the databases are: Celera-PFP, 2,104,820; TSC, 585,811; and Kwok 438,032. Only unique SNPs in the TSC and Kwok data sets were included.

	TSC	Kwok
Celera-PFP	188,694 (0.322)	158,532 (0.362)
TSC		72,024 (0.164)

One means of assessing whether the three sets of SNPs provide the same picture of human variation is to tally the frequencies of the six possible base changes in each set of SNPs (Table 16). Previous measures of nucleotide diversity were mostly derived from small-scale analysis on candidate genes (*101*), and our analysis with all three data sets validates the previous observations at the whole-genome scale. There is remarkable homogeneity between the SNPs found in the Kwok set, the TSC set, and in our whole-genome shotgun (*46*) in this substitution pattern. Compared with the rest of the data sets, Celera-PFP deviates slightly from the 2:1 transition-to-transversion ratio observed in the other SNP sets. This result is not unexpected, because some fraction of the computationally identified SNPs in the Celera-PFP comparison may in fact be sequence errors. A 2:1 transition:transversion ratio for the bona fide SNPs would be obtained if one assumed that 15% of the sequence differences in the Celera-PFP set were a result of (presumably random) sequence errors.

6.3 Estimation of nucleotide diversity from ascertained SNPs

The number of SNPs identified varied widely across chromosomes. In order to normalize these values to the chromosome size and sequence coverage, we used π, the standard statistic for nucleotide diversity (*104*). Nucleotide diversity is a measure of per-site heterozygosity, quantifying the probability that a pair of chromosomes drawn from the population will differ at a nucleotide site. In order to calculate nucleotide diversity for each chromosome, we need to know the number of nucleotide sites that were surveyed for variation, and in methods like reduced respresentation sequencing, we need to know the sequence quality and the depth of coverage at each site. These data are not readily available, so we could not estimate nucleotide diversity from the TSC effort. Estimation of nucleotide diversity from high-quality sequence overlaps should be possible, but again, more information is needed on the details of all the alignments.

Estimation of nucleotide diversity from a shotgun assembly entails calculating for each column of the multialignment, the probability that two or more distinct alleles are present, and the probability of detecting a SNP if in fact the alleles have different sequence (i.e., the probability of correct sequence calls). The greater the depth of coverage and the higher the sequence quality, the higher is the chance of successfully detecting a SNP (*105*). Even after correcting for variation in coverage, the nucleotide diversity appeared to vary across autosomes. The significance of this heterogeneity was tested by analysis of variance, with estimates of π for 100-kbp windows to estimate variability within chromosomes (for the Celera-PFP comparison, $F = 29.73$, $P < 0.0001$).

Average diversity for the autosomes estimated from the Celera-PFP comparison was 8.94×10^{-4}. Nucleotide diversity on the X chromosome was 6.54×10^{-4}. The X is expected to be less variable than autosomes, because for every four copies of autosomes in the population, there are only three X chromosomes, and this smaller effective population size means that random drift will more rapidly remove variation from the X (*106*).

Having ascertained nucleotide variation genome-wide, it appears that previous estimates of nucleotide diversity in humans based on samples of genes were reasonably accurate (*101, 102, 106, 107*). Genome-wide, our estimate of nucleotide diversity was 8.98×10^{-4} for the Celera-PFP alignment, and a published estimate averaged over 10 densely resequenced human genes was 8.00×10^{-4} (*108*).

6.4 Variation in nucleotide diversity across the human genome

Such an apparently high degree of variability among chromosomes in SNP density raises the question of whether there is heterogeneity at a finer scale within chromo-

Table 16. Summary of nucleotide changes in different SNP data sets.

SNP data set	A/G (%)	C/T (%)	A/C (%)	A/T (%)	C/G (%)	T/G (%)	Transition: transversion
Celera-PFP	30.7	30.7	10.3	8.6	9.2	10.3	1.59:1
Kwok*	33.7	33.8	8.5	7.0	8.6	8.4	2.07:1
TSC†	33.3	33.4	8.8	7.3	8.6	8.6	1.99:1

*November 2000 release of the NCBI database dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the method defined as Overlap SnpDetectionWithPolyBayes. The submitter of the data is Pui-Yan Kwok from Washington University. †November 2000 release of NCBI dbSNP (www.ncbi.nlm.nih.gov/SNP/) with the methods defined as TSC-Sanger, TSC-WICGR, and TSC-WUGSC. The submitter of the data is Lincoln Stein from Cold Spring Harbor Laboratory.

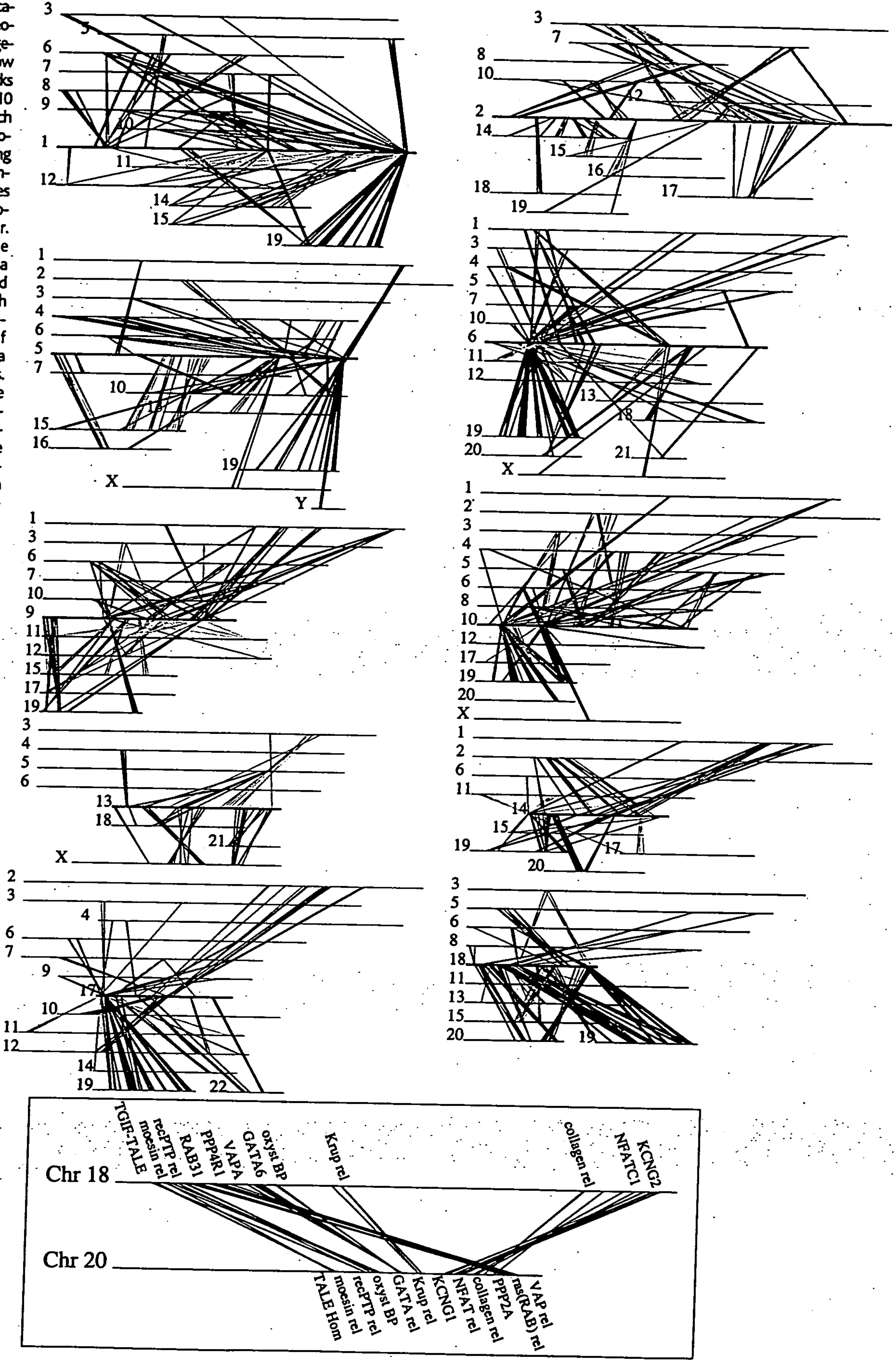
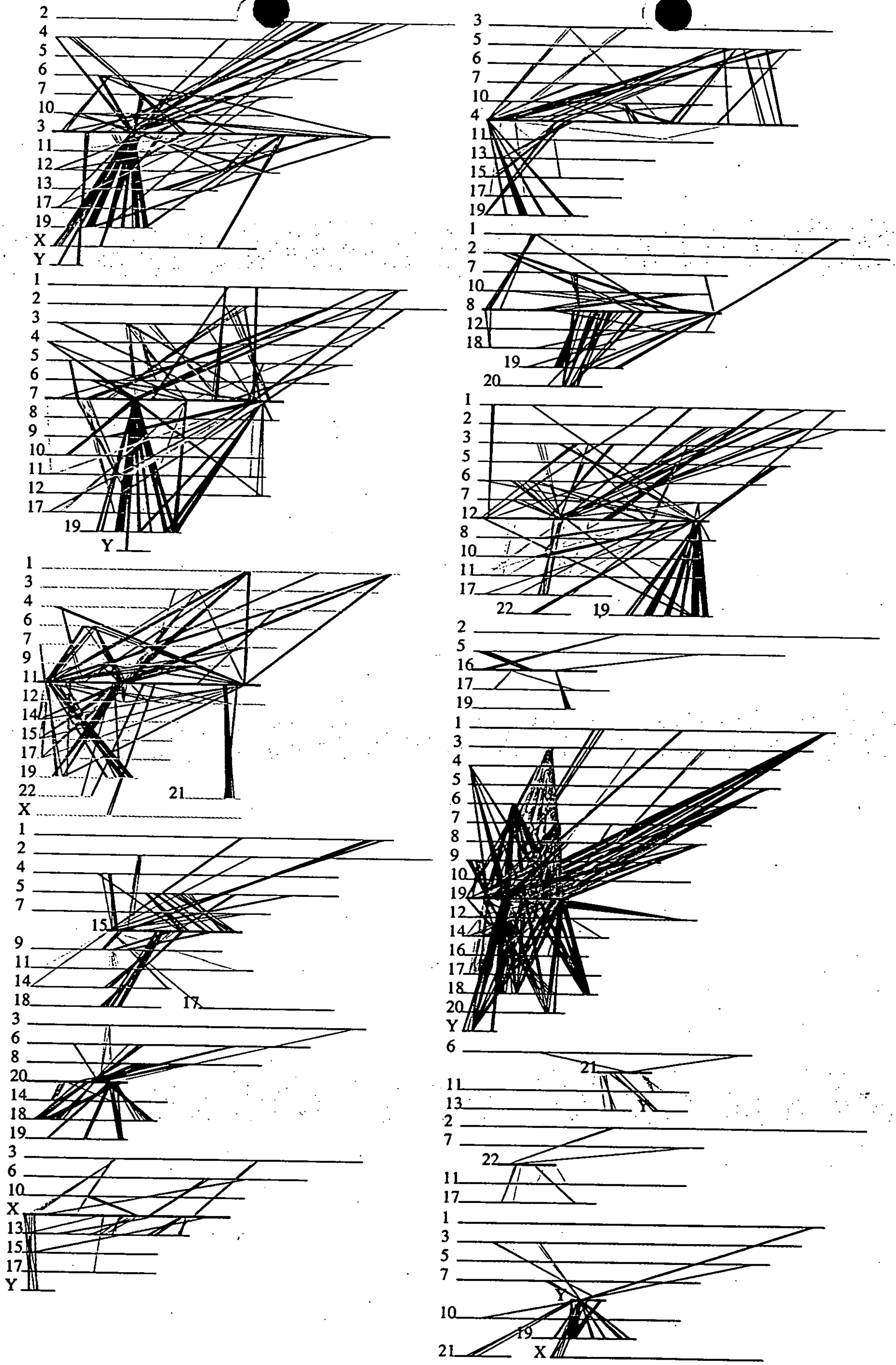


Fig. 13. Segmental duplications between chromosomes in the human genome. The 24 panels show the 1077 duplicated blocks of genes, containing 10,310 pairs of genes in total. Each line represents a pair of homologous genes belonging to a block; all blocks contain at least three genes on each of the chromosomes where they appear. Each panel shows all the duplications between a single chromosome and other chromosomes with shared blocks. The chromosome at the center of each panel is shown as a thick red line for emphasis. Other chromosomes are displayed from top to bottom within each panel ordered by chromosome number. The inset (bottom, center right) shows a close-up of one duplication between chromosomes 18 and 20, expanded to display the gene names of 12 of the 64 gene pairs shown.


2
4
5
6
7
10
3
11
12
13
17
19
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
17
19
Y
1
3
4
6
7
9
11
12
14
15
17
19
22
21
X
1
2
4
5
7
15
9
11
14
18
17
3
6
8
20
14
18
19
3
6
10
X
13
15
17
Y
3
5
6
7
10
4
11
13
15
17
19
1
2
7
10
8
12
18
19
20
1
2
3
5
6
7
12
8
10
11
17
22
19
2
5
16
17
19
1
3
4
5
6
7
8
9
10
19
12
14
16
17
18
20
Y
6
21
11
13
Y
2
7
22
11
17
1
3
5
7
Y
10
19
21
X

somes, and whether this heterogeneity is greater than expected by chance. If SNPs occur by random and independent mutations, then it would seem that there ought to be a Poisson distribution of numbers of SNPs in fragments of arbitrary constant size. The observed dispersion in the distribution of SNPs in 100-kbp fragments was far greater than predicted from a Poisson distribution (Fig. 14). However, this simplistic model ignores the different recombination rates and population histories that exist in different regions of the genome. Population genetics theory holds that we can account for this variation with a mathematical formulation called the neutral coalescent (*109*). Applying well-tested algorithms for simulating the neutral coalescent with recombination (*110*), and using an effective population size of 10,000 and a per-base recombination rate equal to the mutation rate (*111*), we generated a distribution of numbers of SNPs by this model as well (*112*). The observed distribution of SNPs has a much larger variance than either the Poisson model or the coalescent model, and the difference is highly significant. This implies that there is significant variability across the genome in SNP density, an observation that begs an explanation.

Several attributes of the DNA sequence may affect the local density of SNPs, including the rate at which DNA polymerase makes errors and the efficacy of mismatch repair. One key factor that is likely to be associated with SNP density is the G+C content, in part because methylated cytosines in CpG dinucleotides tend to undergo deamination to form thymine, accounting for a nearly 10-fold increase in the mutation rate of CpGs over other dinucleotides. We tallied the GC content and nucleotide diversities in 100-kbp windows across the entire genome and found that the correlation between them was positive ($r = 0.21$) and highly significant ($P < 0.0001$), but G+C content accounted for only a small part of the variation.

6.5 SNPs by genomic class

To test homogeneity of SNP densities across functional classes, we partitioned sites into intergenic (defined as >5 kbp from any predicted transcription unit), 5′-UTR, exonic (missense and silent), intronic, and 3′-UTR for 10,239 known genes, derived from the NCBI RefSeq database and all human genes predicted from the Celera Otto annotation. In coding regions, SNPs were categorized as either silent, for those that do not change amino acid sequence, or missense, for those that change the protein product. The ratio of missense to silent coding SNPs in Celera-PFP, TSC, and Kwok sets (1.12, 0.91, and 0.78, respectively) shows a markedly reduced frequency of missense variants compared with the neutral expectation, consistent with the elimination by natural selection of a fraction of the deleterious amino acid changes (*112*). These ratios are comparable to the missense-to-silent ratios of 0.88 and 1.17 found by Cargill *et al.* (*101*) and by Halushka *et al.* (*102*). Similar results were observed in SNPs derived from Celera shotgun sequences (*46*).

It is striking how small is the fraction of SNPs that lead to potentially dysfunctional alterations in proteins. In the 10,239 RefSeq genes, missense SNPs were only about 0.12, 0.14, and 0.17% of the total SNP counts in Celera-PFP, TSC, and Kwok SNPs, respectively. Nonconservative protein changes constitute an even smaller fraction of missense SNPs (47, 41, and 40% in Celera-PFP, Kwok, and TSC). Intergenic regions have been virtually unstudied (*113*), and we note that 75% of the SNPs we identified were intergenic (Table 17). The SNP rate was highest in introns and lowest in exons. The SNP rate was lower in intergenic regions than in introns, providing one of the first discriminators between these two classes of DNA. These SNP rates were confirmed in the Celera SNPs, which also exhibited a lower rate in exons than in introns, and in extragenic regions than in introns (*46*). Many of these intergenic SNPs will provide valuable information in the form of markers for linkage and association studies, and some fraction is likely to have a regulatory function as well.

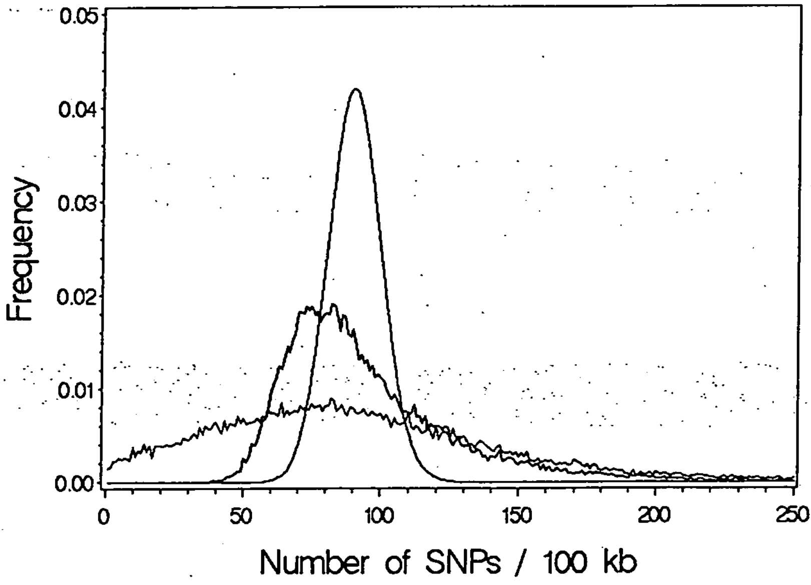


Fig. 14. SNP density in each 100-kbp interval as determined with Celera-PFP SNPs. The color codes are as follows: black, Celera-PFP SNP density; blue, coalescent model; and red, Poisson distribution. The figure shows that the distribution of SNPs along the genome is nonrandom and is not entirely accounted for by a coalescent model of regional history.

7 An Overview of the Predicted Protein-Coding Genes in the Human Genome

Summary. This section provides an initial computational analysis of the predicted protein set with the aim of cataloging prominent differences and similarities when the human genome is compared with other fully sequenced eukaryotic genomes. Over 40% of the predicted protein set in humans cannot be ascribed a molecular function by methods that assign proteins to known families. A protein domain–based analysis provides a detailed catalog of the prominent differences in the human genome when compared with the fly and worm genomes. Prominent among these are domain expansions in proteins involved in developmental regulation and in cellular processes such as neuronal function, hemostasis, acquired immune response, and cytoskeletal complexity. The final enumeration of protein families and details of protein structure will rely on additional experimental work and comprehensive manual curation.

A preliminary analysis of the predicted human protein-coding genes was conducted. Two methods were used to analyze and classify the molecular functions of 26,588 predicted proteins that represent 26,383 gene predictions with at least two lines of evidence as described above. The first method was based on an analysis at the level of protein families, with both the publicly available Pfam database (*114, 115*) and Celera's Panther Classification (CPC) (Fig. 15) (*116*). The second method was based on an analysis at the level of protein domains, with both the Pfam and SMART databases (*115, 117*).

The results presented here are preliminary and are subject to several limitations.

Both the gene predictions and functional assignments have been made by using computational tools, although the statistical models in Panther, Pfam, and SMART have been built, annotated, and reviewed by expert biologists. In the set of computationally predicted genes, we expect both false-positive predictions (some of these may in fact be inactive pseudogenes) and false-negative predictions (some human genes will not be computationally predicted). We also expect errors in delimiting the boundaries of exons and genes. Similarly, in the automatic functional assignments, we also expect both false-positive and false-negative predictions. The functional assignment protocol focuses on protein families that tend to be found across several organisms, or on families of known human genes. Therefore, we do not assign a function to many genes that are not in large families, even if the function is known. Unless otherwise specified, all enumeration of the genes in any given family or functional category was taken from the set of 26,588 predicted proteins, which were assigned functions by using statistical score cutoffs defined for models in Panther, Pfam, and SMART.

For this initial examination of the predicted human protein set, three broad questions were asked: (i) What are the likely molecular functions of the predicted gene products, and how are these proteins categorized with current classification methods? (ii) What are the core functions that appear to be common across the animals? (iii) How does the human protein complement differ from that of other sequenced eukaryotes?

7.1 Molecular functions of predicted human proteins

Figure 15 shows an overview of the putative molecular functions of the predicted 26,588 human proteins that have at least two lines of supporting evidence. About 41% (12,809) of the gene products could not be classified from this initial analysis and are termed proteins with unknown functions. Because our automatic classification methods treat only relatively large protein families, there are a number of "unclassified" sequences that do, in fact, have a known or predicted function. For the 60% of the protein set that have automatic functional predictions, the specific protein functions have been placed into broad classes. We focus here on molecular function (rather than higher order cellular processes) in order to classify as many proteins as possible. These functional predictions are based on similarity to sequences of known function.

In our analysis of the 12,731 additional low-confidence predicted genes (those with only one piece of supporting evidence), only 636 (5%) of these additional putative genes were assigned molecular functions by the automated methods. One-third of these 636 predicted genes represented endogenous retroviral proteins, further suggesting that the majority of these unknown-function genes are not real genes. Given that most of these additional 12,095 genes appear to be unique among the genomes sequenced to date, many may simply represent false-positive gene predictions.

The most common molecular functions are the transcription factors and those involved in nucleic acid metabolism (nucleic acid enzyme). Other functions that are highly represented in the human genome are the receptors, kinases, and hydrolases. Not surprisingly, most of the hydrolases are proteases. There are also many proteins that are members of proto-oncogene families, as well as families of "select regulatory molecules": (i) proteins involved in specific steps of signal transduction such as heterotrimeric GTP-binding proteins (G proteins) and cell cycle regulators, and (ii) proteins that modulate the activity of kinases, G proteins, and phosphatases.

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

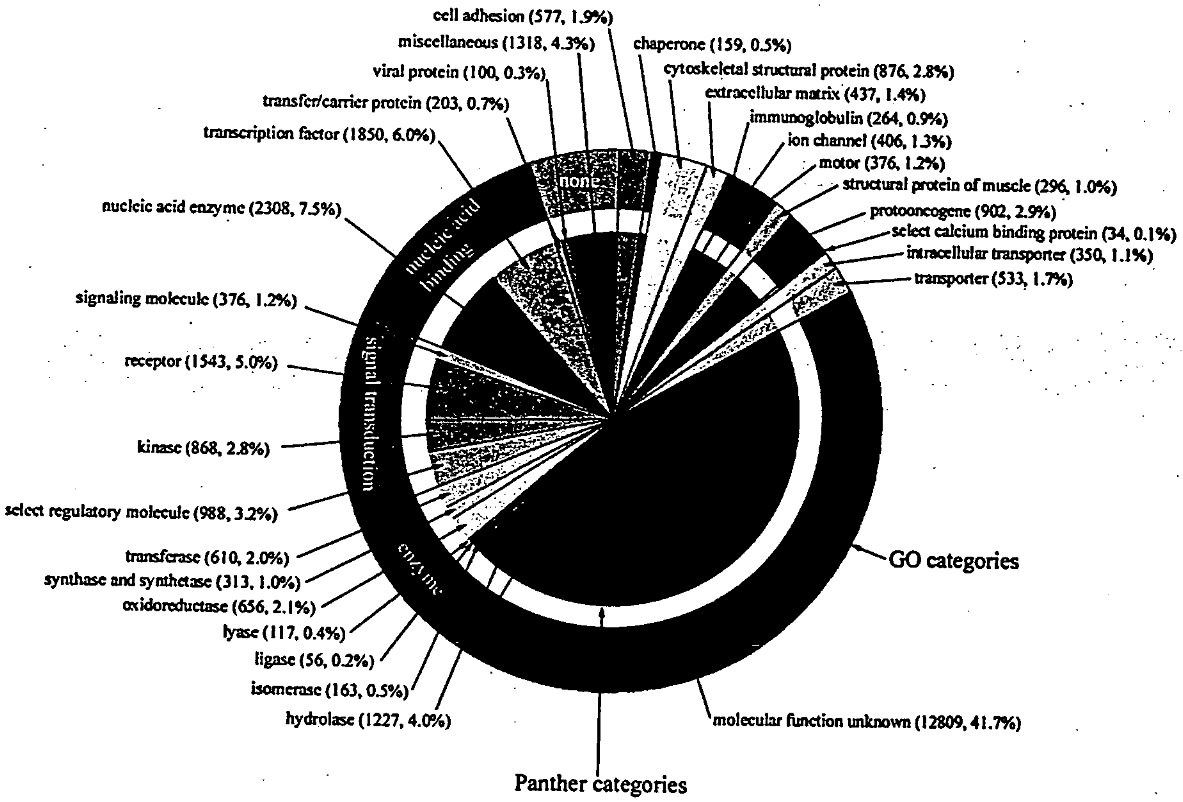


Fig. 15. Distribution of the molecular functions of 26,383 human genes. Each slice lists the numbers and percentages (in parentheses) of human gene functions assigned to a given category of molecular function. The outer circle shows the assignment to molecular function categories in the Gene Ontology (GO) (*179*), and the inner circle shows the assignment to Celera's Panther molecular function categories (*116*).

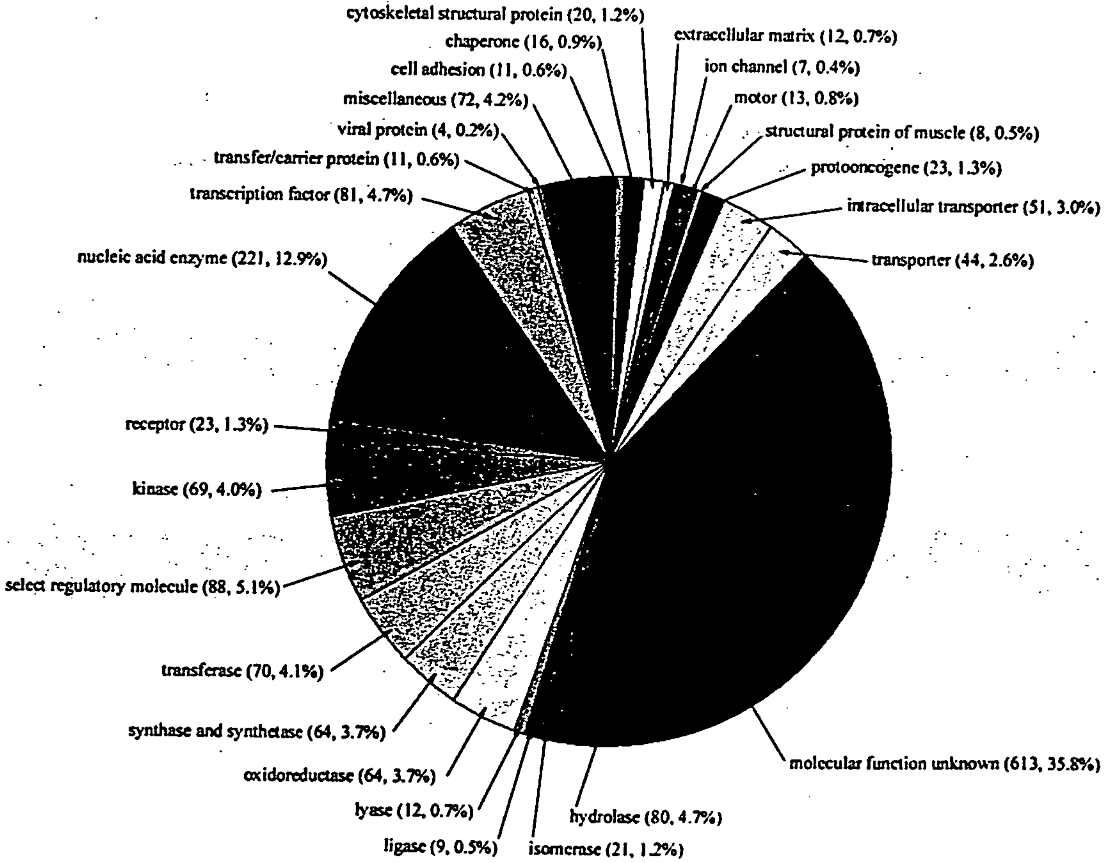
7.2 Evolutionary conservation of core processes

Because of the various "model organism" genome-sequencing projects that have already been completed, reasonable comparative information is available for beginning the analysis of the evolution of the human genome. The genomes of *S. cerevisiae* ("bakers' yeast") (*118*) and two diverse invertebrates, *C. elegans* (a nematode worm) (*119*) and *D. melanogaster* (fly) (*26*), as well as the first plant genome, *A. thaliana*, recently completed (*92*), provide a diverse background for genome comparisons.

We enumerated the "strict orthologs" conserved between human and fly, and between human and worm (Fig. 16) to address the question, What are the core functions that appear to be common across the animals? The concept of orthology is important because if two genes are orthologs, they can be traced by descent to the common ancestor of the two organisms (an "evolutionarily conserved protein set"), and therefore are likely to perform similar conserved functions in the different organisms. It is critical in this analysis to separate orthologs (a gene that appears in two organisms by descent from a common ancestor) from paralogs (a gene that appears in more than one copy in a given organism by a duplication event) because paralogs may subsequently diverge in function. Following the yeast-worm ortholog comparison in (*120*), we identified two different cases for each pairwise comparison (human-fly and human-worm). The first case was a pair of genes, one from each organism, for which there was no other close homolog in either organism. These are straightforwardly identified as orthologous, because there are no additional members of the families that complicate separating orthologs from paralogs. The second case is a family of genes with more than one member in either or both of the organisms being compared. Chervitz *et al.* (*120*) deal with this case by analyzing a phylogenetic tree that described the relationships between all of the sequences in both organisms, and then looked for pairs of genes that were nearest neighbors in the tree. If the nearest-neighbor pairs were from different organisms, those genes were presumed to be orthologs. We note that these nearest neighbors can often be confidently identified from pairwise sequence comparison without having to examine a phylogenetic tree (see legend to Fig. 16). If the nearest neighbors are not from different organisms, there has been a paralogous expansion in one or both organisms after the speciation event (and/or a gene loss by one organism). When this one-to-one correspondence is lost, defining an ortholog becomes ambiguous. For our initial computational overview of the predicted human protein set, we could not answer this question for every predicted protein. Therefore, we consider only "strict orthologs," i.e., the proteins with unambiguous one-to-one relationships (Fig. 16). By these criteria, there are 2758 strict human-fly orthologs, 2031 human-worm (1523 in common between these sets). We define the evolutionarily conserved set as those 1523 human proteins that have strict orthologs in both *D. melanogaster* and *C. elegans*.

The distribution of the functions of the conserved protein set is shown in Fig. 16. Comparison with Fig. 15 shows that, not surprisingly, the set of conserved proteins is not distributed among molecular functions in the same way as the whole human protein set. Compared with the whole human set (Fig. 15), there are several categories that are overrepresented in the conserved set by a factor of ~2 or more. The first category is nucleic acid enzymes, primarily the transcriptional machinery (notably DNA/RNA methyltransferases, DNA/RNA polymerases, helicases, DNA ligases, DNA- and RNA-processing factors, nucleases, and ribosomal proteins). The basic transcriptional and translational machinery is well known to have been conserved over evolution, from bacteria through to the most complex eukaryotes. Many ribonucleoproteins involved in RNA splicing also appear to be conserved among the animals. Other enzyme types are also overrepresented (transferases, oxidoreductases, ligases, lyases, and isomerases). Many of these en-

Fig. 16. Functions of putative orthologs across vertebrate and invertebrate genomes. Each slice lists the number and percentages (in parentheses) of "strict orthologs" between the human, fly, and worm genomes involved in a given category of molecular function. "Strict orthologs" are defined here as bi-directional BLAST best hits (*180*) such that each orthologous pair (i) has a BLASTP *P*-value of $\leq 10^{-10}$ (*120*), and (ii) has a more significant BLASTP score than any paralogs in either organism, i.e., there has likely been no duplication subsequent to speciation that might make the orthology ambiguous. This measure is quite strict and is a lower bound on the number of orthologs. By these criteria, there are 2758 strict human-fly orthologs, and 2031 human-worm orthologs (1523 in common between these sets).



zymes are involved in intermediary metabolism. The only exception is the hydrolase category, which is not significantly overrepresented in the shared protein set. Proteases form the largest part of this category, and several large protease families have expanded in each of these three organisms after their divergence. The category of select regulatory molecules is also overrepresented in the conserved set. The major conserved families are small guanosine triphosphatases (GTPases) (especially the Ras-related superfamily, including ADP ribosylation factor) and cell cycle regulators (particularly the cullin family, cyclin C family, and several cell division protein kinases). The last two significantly overrepresented categories are protein transport and trafficking, and chaperones. The most conserved groups in these categories are proteins involved in coated vesicle-mediated transport, and chaperones involved in protein folding and heat-shock response [particularly the DNAJ family, and heat-shock protein 60 (HSP60), HSP70, and HSP90 families]. These observations provide only a conservative estimate of the protein families in the context of specific cellular processes that were likely derived from the last common ancestor of the human, fly, and worm. As stated before, this analysis does not provide a complete estimate of conservation across the three animal genomes, as paralogous duplication makes the determination of true orthologs difficult within the members of conserved protein families.

7.3 Differences between the human genome and other sequenced eukaryotic genomes

To explore the molecular building blocks of the vertebrate taxon, we have compared the human genome with the other sequenced eukaryotic genomes at three levels: molecular functions, protein families, and protein domains.

Molecular differences can be correlated with phenotypic differences to begin to reveal the developmental and cellular processes that are unique to the vertebrates. Tables 18 and 19 display a comparison among all sequenced eukaryotic genomes, over selected protein/domain families (defined by sequence similarity, e.g., the serine-threonine protein kinases) and superfamilies (defined by shared molecular function, which may include several sequence-related families, e.g., the cytokines). In these tables we have focused on (super) families that are either very large or that differ significantly in humans compared with the other sequenced eukaryote genomes. We have found that the most prominent human expansions are in proteins involved in (i) acquired immune functions; (ii) neural development, structure, and functions; (iii) intercellular and intracellular signaling pathways in development and homeostasis; (iv) hemostasis; and (v) apoptosis.

Acquired immunity. One of the most striking differences between the human genome and the *Drosophila* or *C. elegans* genome is the appearance of genes involved in acquired immunity (Tables 18 and 19). This is expected, because the acquired immune response is a defense system that only occurs in vertebrates. We observe 22 class I and 22 class II major histocompatibility complex (MHC) antigen genes and 114 other immunoglobulin genes in the human genome. In addition, there are 59 genes in the cognate immunoglobulin receptor family. At the domain level, this is exemplified by an expansion and recruitment of the ancient immunoglobulin fold to constitute molecules such as MHC, and of the integrin fold to form several of the cell adhesion molecules that mediate interactions between immune effector cells and the extracellular matrix. Vertebrate-specific proteins include the paracrine immune regulators family of secreted 4-alpha helical bundle proteins, namely the cytokines and chemokines. Some of the cytoplasmic signal transduction components associated with cytokine receptor signal transduction are also features that are poorly represented in the fly and worm. These include protein domains found in the signal transducer and activator of transcription (STATs), the suppressors of cytokine signaling (SOCS), and protein inhibitors of activated STATs (PIAS). In contrast, many of the animal-specific protein domains that play a role in innate immune response, such as the Toll receptors, do not appear to be significantly expanded in the human genome.

Neural development, structure, and function. In the human genome, as compared with the worm and fly genomes, there is a marked increase in the number of members of protein families that are involved in neural development. Examples include neurotrophic factors such as ependymin, nerve growth factor, and signaling molecules such as semaphorins, as well as the number of proteins involved directly in neural structure and function such as myelin proteins, voltage-gated ion channels, and synaptic proteins such as synaptotagmin. These observations correlate well with the known phenotypic differences between the nervous systems of these taxa, notably (i) the increase in the number and connectivity of neurons; (ii) the increase in number of distinct neural cell types (as many as a thousand or more in human compared with a few hundred in fly and worm) (*121*); (iii) the increased length of individual axons; and (iv) the significant increase in glial cell number, especially the appearance of myelinating glial cells, which are electrically inert supporting cells differentiated from the same stem cells as neurons. A number of prominent protein expansions are involved in the processes of neural development. Of the extracellular domains that mediate cell adhesion, the connexin domain–containing proteins (*122*) exist only in humans. These proteins, which are not present in the *Drosophila* or *C. elegans* genomes, appear to provide the constitutive subunits of intercellular channels and the structural basis for electrical coupling. Pathway finding by axons and neuronal network formation is mediated through a subset of ephrins and their cognate receptor tyrosine kinases that act as positional labels to establish topographical projections (*123*). The probable biological role for the semaphorins (22 in human compared with 6 in the fly and 2 in the worm) and their receptors (neuropilins and plexins) is that of axonal guidance molecules (*124*). Signaling molecules such as neurotrophic factors and some cytokines have been shown to regulate neuronal cell survival, proliferation, and axon guidance (*125*). Notch receptors and ligands play important roles in glial cell fate determination and gliogenesis (*126*).

Other human expanded gene families play key roles directly in neural structure and function. One example is synaptotagmin (expanded more than twofold in humans relative to the invertebrates), originally found to regulate synaptic transmission by serving as a Ca^{2+} sensor (or receptor) during synaptic vesicle fusion and release (*127*). Of interest is the increased co-occurrence in humans of PDZ and the SH3 domains in neuronal-specific adaptor molecules; examples include proteins that likely modulate channel activity at synaptic junctions (*128*). We also noted expansions in several ion-channel families (Table 19), including the EAG subfamily (related to cyclic nucleotide gated channels), the voltage-gated calcium/sodium channel family, the inward-rectifier potassium channel family, and the voltage-gated potassium channel, alpha subunit family. Voltage-gated sodium and potassium channels are involved in the generation of action potentials in neurons. Together with voltage-gated calcium channels, they also play a key role in coupling action potentials to neurotransmitter release, in the development of neurites, and in short-term memory. The recent observation of a calcium-regulated association between sodium channels and synaptotagmin may have consequences for the establishment and regulation of neuronal excitability (*129*).

Myelin basic protein and myelin-associated glycoprotein are major classes of protein components in both the central and peripheral nervous system of vertebrates. Myelin P0 is a major component of peripheral myelin, and myelin proteolipid and myelin oligodendrocyte glycopotein are found in the central nervous system. Mutations in any of these

Table 18. Domain-based comparative analysis of proteins in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A). The predicted protein set of each of the above eukaryotic organisms was analyzed with Pfam version 5.5 using E value cutoffs of 0.001. The number of proteins containing the specified Pfam domains as well as the total number of domains (in parentheses) are shown in each column. Domains were categorized into cellular processes for presentation. Some domains (i.e., SH2) are listed in more than one cellular process. Results of the Pfam analysis may differ from results obtained based on human curation of protein families, owing to the limitations of large-scale automatic classifications. Representative examples of domains with reduced counts owing to the stringent E value cutoff used for this analysis are marked with a double asterisk (**). Examples include short divergent and predominantly alpha-helical domains, and certain classes of cysteine-rich zinc finger proteins.

Accession number	Domain name	Domain description	H	F	W	Y	A
		Developmental and homeostatic regulators					
PF02039	Adrenomedullin	Adrenomedullin	1	0	0	0	0
PF00212	ANP	Atrial natriuretic peptide	2	0	0	0	0
PF00028	Cadherin	Cadherin domain	100 (550)	14 (157)	16 (66)	0	0
PF00214	Calc_CGRP_IAPP	Calcitonin/CGRP/IAPP family	3	0	0	0	0
PF01110	CNTF	Ciliary neurotrophic factor	1	0	0	0	0
PF01093	Clusterin	Clusterin	3	0	0	0	0
PF00029	Connexin	Connexin	14 (16)	0	0	0	0
PF00976	ACTH_domain	Corticotropin ACTH domain	1	0	0	0	0
PF00473	CRF	Corticotropin-releasing factor family	2	1	0	0	0
PF00007	Cys_knot	Cystine-knot domain	10 (11)	2	0	0	0
PF00778	DIX	Dix domain	5	2	4	0	0
PF00322	Endothelin	Endothelin family	3	0	0	0	0
PF00812	Ephrin	Ephrin	7 (8)	2	4	0	0
PF01404	EPh_lbd	Ephrin receptor ligand binding domain	12	2	1	0	0
PF00167	FGF	Fibroblast growth factor	23	1	1	0	0
PF01534	Frizzled	Frizzled/Smoothened family membrane region	9	7	3	0	0
PF00236	Hormone6	Glycoprotein hormones	1	0	0	0	0
PF01153	Glypican	Glypican	14	2	1	0	0
PF01271	Granin	Grainin (chromogranin or secretogranin)	3	0	0	0	0
PF02058	Guanylin	Guanylin precursor	1	0	0	0	0
PF00049	Insulin	Insulin/IGF/Relaxin family	7	4	0	0	0
PF00219	IGFBP	Insulin-like growth factor binding proteins	10	0	0	0	0
PF02024	Leptin	Leptin	1	0	0	0	0
PF00193	Xlink	LINK (hyaluron binding)	13 (23)	0	1	0	0
PF00243	NGF	Nerve growth factor family	3	0	0	0	0
PF02158	Neuregulin	Neuregulin family	4	0	0	0	0
PF00184	Hormone5	Neurohypophysial hormones	1	0	0	0	0
PF02070	NMU	Neuromedin U	1	0	0	0	0
PF00066	Notch	Notch (DSL) domain	3 (5)	2 (4)	2 (6)	0	0
PF00865	Osteopontin	Osteopontin	1	0	0	0	0
PF00159	Hormone3	Pancreatic hormone peptides	3	0	0	0	0
PF01279	Parathyroid	Parathyroid hormone family	2	0	0	0	0
PF00123	Hormone2	Peptide hormone	5 (9)	0	0	0	0
PF00341	PDGF	Platelet-derived growth factor (PDGF)	5	1	0	0	0
PF01403	Sema	Sema domain	27 (29)	8 (10)	3 (4)	0	0
PF01033	Somatomedin_B	Somatomedin B domain	5 (8)	3	0	0	0
PF00103	Hormone	Somatotropin	1	0	0	0	0
PF02208	Sorb	Sorbin homologous domain	2	0	0	0	0
PF02404	SCF	Stem cell factor	2	0	0	0	0
PF01034	Syndecan	Syndecan domain	3	1	1	0	0
PF00020	TNFR_c6	TNFR/NGFR cysteine-rich region	17 (31)	1	0	0	0
PF00019	TGF-β	Transforming growth factor β-like domain	27 (28)	6	4	0	0
PF01099	Uteroglobin	Uteroglobin family	3	0	0	0	0
PF01160	Opiods_neuropep	Vertebrate endogenous opioids neuropeptide	3	0	0	0	0
PF00110	Wnt	Wnt family of developmental signaling proteins	18	7 (10)	5	0	0
		Hemostasis					
PF01821	ANATO	Anaphylotoxin-like domain	6 (14)	0	0	0	0
PF00386	C1q	C1q domain	24	0	0	0	0
PF00200	Disintegrin	Disintegrin	18	2	3	0	0
PF00754	F5_F8_type_C	F5/8 type C domain	15 (20)	5 (6)	2	0	0
PF01410	COLFI	Fibrillar collagen C-terminal domain	10	0	0	0	0
PF00039	Fn1	Fibronectin type I domain	5 (18)	0	0	0	0
PF00040	Fn2	Fibronectin type II domain	11 (16)	0	0	0	0
PF00051	Kringle	Kringle domain	15 (24)	2	2	0	0
PF01823	MACPF	MAC/Perforin domain	6	0	0	0	0
PF00354	Pentaxin	Pentaxin family	9	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF02210	TSPN	Thrombospondin N-terminal–like domains	14	1	0	0	0
PF01108	Tissue_fac	Tissue factor	1	0	0	0	0
PF00868	Transglutamin_N	Transglutaminase family	6	1	0	0	0
PF00927	Transglutamin_C	Transglutaminase family	8	1	0	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00594	Gla	Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain	11	0	0	0	0
		Immune response					
PF00711	Defensin_beta	Beta defensin	1	0	0	0	0
PF00748	Calpain_inhib	Calpain inhibitor repeat	3 (9)	0	0	0	0
PF00666	Cathelicidins	Cathelicidins	2	0	0	0	0
PF00129	MHC_I	Class I histocompatibility antigen, domains alpha 1 and 2	18 (20)	0	0	0	0
PF00993	MHC_II_alpha**	Class II histocompatibility antigen, alpha domain	5 (6)	0	0	0	0
PF00969	MHC_II_beta**	Class II histocompatibility antigen, beta domain	7	0	0	0	0
PF00879	Defensin_propep	Defensin propeptide	3	0	0	0	0
PF01109	GM_CSF	Granulocyte-macrophage colony-stimulating factor	1	0	0	0	0
PF00047	Ig	Immunoglobulin domain	381 (930)	125 (291)	67 (323)	0	0
PF00143	Interferon	Interferon alpha/beta domain	7 (9)	0	0	0	0
PF00714	IFN-gamma	Interferon gamma	1	0	0	0	0
PF00726	IL10	Interleukin-10	1	0	0	0	0
PF02372	IL15	Interleukin-15	1	0	0	0	0
PF00715	IL2	Interleukin-2	1	0	0	0	0
PF00727	IL4	Interleukin-4	1	0	0	0	0
PF02025	IL5	Interleukin-5	1	0	0	0	0
PF01415	IL7	Interleukin-7/9 family	1	0	0	0	0
PF00340	IL1	Interleukin-1	7	0	0	0	0
PF02394	IL1_propep	Interleukin-1 propeptide	1	0	0	0	0
PF02059	IL3	Interleukin-3	1	0	0	0	0
PF00489	IL6	Interleukin-6/G-CSF/MGF family	2	0	0	0	0
PF01291	LIF_OSM	Leukemia inhibitory factor (LIF)/oncostatin (OSM) family	2	0	0	0	0
PF00323	Defensins	Mammalian defensin	2	0	0	0	0
PF01091	PTN_MK	PTN/MK heparin-binding protein	2	0	0	0	0
PF00277	SAA_proteins	Serum amyloid A protein	4	0	0	0	0
PF00048	IL8	Small cytokines (intecrine/chemokine), interleukin-8 like	32	0	0	0	0
PF01582	TIR	TIR domain	18	8	2	0	131 (143)
PF00229	TNF	TNF (tumor necrosis factor) family	12	0	0	0	0
PF00088	Trefoil	Trefoil (P-type) domain	5 (6)	0	2	0	0
		PI-PY-rho GTPase signaling					
PF00779	BTK	BTK motif	5	1	0	0	0
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00609	DAGKa	Diacylglycerol kinase accessory domain (presumed)	9	4	7	0	6
PF00781	DAGKc	Diacylglycerol kinase catalytic domain (presumed)	10	8	8	2	11 (12)
PF00610	DEP	Domain found in Dishevelled, Egl-10, and Pleckstrin (DEP)	12 (13)	4	10	5	2
PF01363	FYVE	FYVE zinc finger	28 (30)	14	15	5	15
PF00996	GDI	GDP dissociation inhibitor	6	2	1	1	3
PF00503	G-alpha	G-protein alpha subunit	27 (30)	10	20 (23)	2	5
PF00631	G-gamma	G-protein gamma like domains	16	5	5	1	0
PF00616	RasGAP	GTPase-activator protein for Ras-like GTPase	11	5	8	3	0
PF00618	RasGEFN	Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif	9	2	3	5	0
PF00625	Guanylate_kin	Guanylate kinase	12	8	7	1	4
PF02189	ITAM	Immunoreceptor tyrosine-based activation motif	3	0	0	0	0
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF00130	DAG_PE-bind	Phorbol esters/diacylglycerol binding domain (C1 domain)	45 (56)	25 (31)	26 (40)	1 (2)	4
PF00388	PI-PLC-X	Phosphatidylinositol-specific phospholipase C, X domain	12	3	7	1	8
PF00387	PI-PLC-Y	Phosphatidylinositol-specific phospholipase C, Y domain	11	2	7	1	8
PF00640	PID	Phosphotyrosine interaction domain (PTB/PID)	24 (27)	13	11 (12)	0	0
PF02192	PI3K_p85B	PI3-kinase family, p85-binding domain	2	1	1	0	0
PF00794	PI3K_rbd	PI3-kinase family, ras-binding domain	6	3	1	0	0
PF01412	ArfGAP	Putative GTP-ase activating protein for Arf	16	9	8	6	15
PF02196	RBD	Raf-like Ras-binding domain	6 (7)	4	1	0	0
PF02145	Rap_GAP	Rap/ran-GAP	5	4	2	0	0
PF00788	RA	Ras association (RalGDS/AF-6) domain	18 (19)	7 (9)	6	1	0
PF00071	Ras	Ras family	126	56 (57)	51	23	78
PF00617	RasGEF	RasGEF domain	21	8	7	5	0
PF00615	RGS	Regulator of G protein signaling domain	27	6 (7)	12 (13)	1	0
PF02197	RIIa	Regulatory subunit of type II PKA R-subunit	4	1	2	1	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00620	RhoGAP	RhoGAP domain	59	19	20	9	8
PF00621	RhoGEF	RhoGEF domain	46	23 (24)	18 (19)	3	0
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01017	STAT	STAT protein	7	1	1 (2)	0	0
PF00790	VHS	VHS domain	4	2	4	4	8
PF00568	WH1	WH1 domain	7	2	2 (3)	1	0
		Domains involved in apoptosis					
PF00452	Bcl-2	Bcl-2	9	2	1	0	0
PF02180	BH4	Bcl-2 homology region 4	3	0	1	0	0
PF00619	CARD	Caspase recruitment domain	16	0	2	0	0
PF00531	Death	Death domain	16	5	7	0	0
PF01335	DED	Death effector domain	4 (5)	0	0	0	0
PF02179	BAG	Domain present in Hsp70 regulators	5 (8)	3	2	1	5
PF00656	ICE_p20	ICE-like protease (caspase) p20 domain	11	7	3	0	0
PF00653	BIR	Inhibitor of Apoptosis domain	8 (14)	5 (9)	2 (3)	1 (2)	0
		Cytoskeletal					
PF00022	Actin	Actin	61 (64)	15 (16)	12	9 (11)	24
PF00191	Annexin	Annexin	16 (55)	4 (16)	4 (11)	0	6 (16)
PF00402	Calponin	Calponin family	13 (22)	3	7 (19)	0	0
PF00373	Band_41	FERM domain (Band 4.1 family)	29 (30)	17 (19)	11 (14)	0	0
PF00880	Nebulin_repeat	Nebulin repeat	4 (148)	1 (2)	1	0	0
PF00681	Plectin_repeat	Plectin repeat	2 (11)	0	0	0	0
PF00435	Spectrin	Spectrin repeat	31 (195)	13 (171)	10 (93)	0	0
PF00418	Tubulin-binding	Tau and MAP proteins, tubulin-binding	4 (12)	1 (4)	2 (8)	0	0
PF00992	Troponin	Troponin	4	6	8	0	0
PF02209	VHP	Villin headpiece domain	5	2	2	0	5
PF01044	Vinculin	Vinculin family	4	2	1	0	0
		ECM adhesion					
PF01391	Collagen	Collagen triple helix repeat (20 copies)	65 (279)	10 (46)	174 (384)	0	0
PF01413	C4	C-terminal tandem repeated domain in type 4 procollagen	6 (11)	2 (4)	3 (6)	0	0
PF00431	CUB	CUB domain	47 (69)	9 (47)	43 (67)	0	0
PF00008	EGF	EGF-like domain	108 (420)	45 (186)	54 (157)	0	1
PF00147	Fibrinogen_C	Fibrinogen beta and gamma chains, C-terminal globular domain	26	10 (11)	6	0	0
PF00041	Fn3	Fibronectin type III domain	106 (545)	42 (168)	34 (156)	0	1
PF00757	Furin-like	Furin-like cysteine rich region	5	2	1	0	0
PF00357	Integrin_A	Integrin alpha cytoplasmic region	3	1	2	0	0
PF00362	Integrin_B	Integrins, beta chain	8	2	2	0	0
PF00052	Laminin_B	Laminin B (Domain IV)	8 (12)	4 (7)	6 (10)	0	0
PF00053	Laminin_EGF	Laminin EGF-like (Domains III and V)	24 (126)	9 (62)	11 (65)	0	0
PF00054	Laminin_G	Laminin G domain	30 (57)	18 (42)	14 (26)	0	0
PF00055	Laminin_Nterm	Laminin N-terminal (Domain VI)	10	6	4	0	0
PF00059	Lectin_c	Lectin C-type domain	47 (76)	23 (24)	91 (132)	0	0
PF01463	LRRCT	Leucine rich repeat C-terminal domain	69 (81)	23 (30)	7 (9)	0	0
PF01462	LRRNT	Leucine rich repeat N-terminal domain	40 (44)	7 (13)	3 (6)	0	0
PF00057	Ldl_recept_a	Low-density lipoprotein receptor domain class A	35 (127)	33 (152)	27 (113)	0	0
PF00058	Ldl_recept_b	Low-density lipoprotein receptor repeat class B	15 (96)	9 (56)	7 (22)	0	0
PF00530	SRCR	Scavenger receptor cysteine-rich domain	11 (46)	4 (8)	1 (2)	0	0
PF00084	Sushi	Sushi domain (SCR repeat)	53 (191)	11 (42)	8 (45)	0	0
PF00090	Tsp_1	Thrombospondin type 1 domain	41 (66)	11 (23)	18 (47)	0	0
PF00092	Vwa	von Willebrand factor type A domain	34 (58)	0	17 (19)	0	1
PF00093	Vwc	von Willebrand factor type C domain	19 (28)	6 (11)	2 (5)	0	0
PF00094	Vwd	von Willebrand factor type D domain	15 (35)	3 (7)	9	0	0
		Protein interaction domains					
PF00244	14-3-3	14-3-3 proteins	20	3	3	2	15
PF00023	Ank	Ank repeat	145 (404)	72 (269)	75 (223)	12 (20)	66 (111)
PF00514	Armadillo_seg	Armadillo/beta-catenin-like repeats	22 (56)	11 (38)	3 (11)	2 (10)	25 (67)
PF00168	C2	C2 domain	73 (101)	32 (44)	24 (35)	6 (9)	66 (90)
PF00027	cNMP_binding	Cyclic nucleotide-binding domain	26 (31)	21 (33)	15 (20)	2 (3)	22
PF01556	DnaJ_C	DnaJ C terminal region	12	9	5	3	19
PF00226	DnaJ	DnaJ domain	44	34	33	20	93
PF00036	Efhand**	EF hand	83 (151)	64 (117)	41 (86)	4 (11)	120 (328)
PF00611	FCH	Fes/CIP4 homology domain	9	3	2	4	0
PF01846	FF	FF domain	4 (11)	4 (10)	3 (16)	2 (5)	4 (8)
PF00498	FHA	FHA domain	13	15	7	13 (14)	17

myelin proteins result in severe demyelination, which is a pathological condition in which the myelin is lost and the nerve conduction is severely impaired (*130*). Humans have at least 10 genes belonging to four different families involved in myelin production (five myelin P0, three myelin proteolipid, myelin basic protein, and myelin-oligodendrocyte glycoprotein, or MOG), and possibly more-remotely related members of the MOG family. Flies have only a single myelin proteolipid, and worms have none at all.

Intercellular and intracellular signaling pathways in development and homeostasis. Many protein families that have expanded in humans relative to the invertebrates are involved in signaling processes, particularly in response to development and differentiation

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF00254	FKBP	FKBP-type peptidyl-prolyl cis-trans isomerases	15 (20)	7 (8)	7 (13)	4	24 (29)
PF01590	GAF	GAF domain	7 (8)	2 (4)	1	0	10
PF01344	Kelch	Kelch motif	54 (157)	12 (48)	13 (41)	3	102 (178)
PF00560	LRR**	Leucine Rich Repeat	25 (30)	24 (30)	7 (11)	1	15 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00989	PAS	PAS domain	18 (19)	9 (10)	6	1	13 (18)
PF00595	PDZ	PDZ domain (Also known as DHR or GLGF)	96 (154)	60 (87)	46 (66)	2	5
PF00169	PH	PH domain	193 (212)	72 (78)	65 (68)	24	23
PF01535	PPR**	PPR repeat	5	3 (4)	0	1	474 (2485)
PF00536	SAM	SAM domain (Sterile alpha motif)	29 (31)	15	8	3	6
PF01369	Sec7	Sec7 domain	13	5	5	5	9
PF00017	SH2	Src homology 2 (SH2) domain	87 (95)	33 (39)	44 (48)	1	3
PF00018	SH3	Src homology 3 (SH3) domain	143 (182)	55 (75)	46 (61)	23 (27)	4
PF01740	STAS	STAS domain	5	1	6	2	13
PF00515	TPR**	TPR domain	72 (131)	39 (101)	28 (54)	16 (31)	65 (124)
PF00400	WD40**	WD40 domain	136 (305)	98 (226)	72 (153)	56 (121)	167 (344)
PF00397	WW	WW domain	32 (53)	24 (39)	16 (24)	5 (8)	11 (15)
PF00569	ZZ	ZZ-Zinc finger present in dystrophin, CBP/p300	10 (11)	13	10	2	10
		Nuclear interaction domains					
PF01754	Zf-A20	A20-like zinc finger	2 (8)	2	2	0	8
PF01388	ARID	ARID DNA binding domain	11	6	4	2	7
PF01426	BAH	BAH domain	8 (10)	7 (8)	4 (5)	5	21 (25)
PF00643	Zf-B_box**	B-box zinc finger	32 (35)	1	2	0	0
PF00533	BRCT	BRCA1 C Terminus (BRCT) domain	17 (28)	10 (18)	23 (35)	10 (16)	12 (16)
PF00439	Bromodomain	Bromodomain	37 (48)	16 (22)	18 (26)	10 (15)	28
PF00651	BTB	BTB/POZ domain	97 (98)	62 (64)	86 (91)	1 (2)	30 (31)
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	3 (4)	1	0	0	13 (15)
PF00385	Chromo	chromo' (CHRromatin Organization MOdifier) domain	24 (27)	14 (15)	17 (18)	1 (2)	12
PF00125	Histone	Core histone H2A/H2B/H3/H4	75 (81)	5	71 (73)	8	48
PF00134	Cyclin	Cyclin	19	10	10	11	35
PF00270	DEAD	DEAD/DEAH box helicase	63 (66)	48 (50)	55 (57)	50 (52)	84 (87)
PF01529	Zf-DHHC	DHHC zinc finger domain	15	20	16	7	22
PF00646	F-box**	F-box domain	16	15	309 (324)	9	165 (167)
PF00250	Fork_head	Fork head domain	35 (36)	20 (21)	15	4	0
PF00320	GATA	GATA zinc finger	11 (17)	5(6)	8 (10)	9	26
PF01585	G-patch	G-patch domain	18	16	13	4	14 (15)
PF00010	HLH**	Helix-loop-helix DNA-binding domain	60 (61)	44	24	4	39
PF00850	Hist_deacetyl	Histone deacetylase family	12	5 (6)	8 (10)	5	10
PF00046	Homeobox	Homeobox domain	160 (178)	100 (103)	82 (84)	6	66
PF01833	TIG	IPT/TIG domain	29 (53)	11 (13)	5 (7)	2	1
PF02373	JmjC	JmjC domain	10	4	6	4	7
PF02375	JmjN	JmjN domain	7	4	2	3	7
PF00013	KH-domain	KH domain	28 (67)	14 (32)	17 (46)	4 (14)	27 (61)
PF01352	KRAB	KRAB box	204 (243)	0	0	0	0
PF00104	Hormone_rec	Ligand-binding domain of nuclear hormone receptor	47	17	142 (147)	0	0
PF00412	LIM	LIM domain containing proteins	62 (129)	33 (83)	33 (79)	4 (7)	10 (16)
PF00917	MATH	MATH domain	11	5	88 (161)	1	61 (74)
PF00249	Myb_DNA-binding	Myb-like DNA-binding domain	32 (43)	18 (24)	17 (24)	15 (20)	243 (401)
PF02344	Myc-LZ	Myc leucine zipper domain	1	0	0	0	0
PF01753	Zf-MYND	MYND finger	14	14	9	1	7
PF00628	PHD	PHD-finger	68 (86)	40 (53)	32 (44)	14 (15)	96 (105)
PF00157	Pou	Pou domain—N-terminal to homeobox domain	15	5	4	0	0
PF02257	RFX_DNA_binding	RFX DNA-binding domain	7	2	1	1	0
PF00076	Rrm	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	224 (324)	127 (199)	94 (145)	43 (73)	232 (369)
PF02037	SAP	SAP domain	15	8	5	5	6 (7)
PF00622	SPRY	SPRY domain	44 (51)	10 (12)	5 (7)	3	6
PF01852	START	START domain	10	2	6	0	23
PF00907	T-box	T-box	17 (19)	8	22	0	0

Table 18 (*Continued*)

Accession number	Domain name	Domain description	H	F	W	Y	A
PF02135	Zf-TAZ	TAZ finger	2 (3)	1 (2)	6 (7)	0	10 (15)
PF01285	TEA	TEA domain	4	1	1	1	0
PF02176	Zf-TRAF	TRAF-type zinc finger	6 (9)	1 (3)	1	0	2
PF00352	TBP	Transcription factor TFIID (or TATA-binding protein, TBP)	2 (4)	4 (8)	2 (4)	1 (2)	2 (4)
PF00567	TUDOR	TUDOR domain	9 (24)	9 (19)	4 (5)	0	2
PF00642	Zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	17 (22)	6 (8)	22 (42)	3 (5)	31 (46)
PF00096	Zf-C2H2**	ZInc finger, C2H2 type	564 (4500)	234 (771)	68 (155)	34 (56)	21 (24)
PF00097	Zf-C3HC4	Zinc finger, C3HC4 type (RING finger)	135 (137)	57	88 (89)	18	298 (304)
PF00098	Zf-CCHC	Zinc knuckle	9 (17)	6 (10)	17 (33)	7 (13)	68 (91)

(Tables 18 and 19). They include secreted hormones and growth factors, receptors, intracellular signaling molecules, and transcription factors.

Developmental signaling molecules that are enriched in the human genome include growth factors such as wnt, transforming growth factor–β (TGF-β), fibroblast growth factor (FGF), nerve growth factor, platelet derived growth factor (PDGF), and ephrins. These growth factors affect tissue differentiation and a wide range of cellular processes involving actin-cytoskeletal and nuclear regulation. The corresponding receptors of these developmental ligands are also expanded in humans. For example, our analysis suggests at least 8 human ephrin genes (2 in the fly, 4 in the worm) and 12 ephrin receptors (2 in the fly, 1 in the worm). In the wnt signaling pathway, we find 18 wnt family genes (6 in the fly, 5 in the worm) and 12 frizzled receptors (6 in the fly, 5 in the worm). The Groucho family of transcriptional corepressors downstream in the wnt pathway are even more markedly expanded, with 13 predicted members in humans (2 in the fly, 1 in the worm).

Extracellular adhesion molecules involved in signaling are expanded in the human genome (Tables 18 and 19). The interactions of several of these adhesion domains with extracellular matrix proteoglycans play a critical role in host defense, morphogenesis, and tissue repair (*131*). Consistent with the well-defined role of heparan sulfate proteoglycans in modulating these interactions (*132*), we observe an expansion of the heparin sulfate sulfotransferases in the human genome relative to worm and fly. These sulfotransferases modulate tissue differentiation (*133*). A similar expansion in humans is noted in structural proteins that constitute the actin-cytoskeletal architecture. Compared with the fly and worm, we observe an explosive expansion of the nebulin (35 domains per protein on average), aggrecan (12 domains per protein on average), and plectin (5 domains per protein on average) repeats in humans. These repeats are present in proteins involved in modulating the actin-cytoskeleton with predominant expression in neuronal, muscle, and vascular tissues.

Comparison across the five sequenced eukaryotic organisms revealed several expanded protein families and domains involved in cytoplasmic signal transduction (Table 18). In particular, signal transduction pathways playing roles in developmental regulation and acquired immunity were substantially enriched. There is a factor of 2 or greater expansion in humans in the Ras superfamily GTPases and the GTPase activator and GTP exchange factors associated with them. Although there are about the same number of tyrosine kinases in the human and *C. elegans* genomes, in humans there is an increase in the SH2, PTB, and ITAM domains involved in phosphotyrosine signal transduction. Further, there is a twofold expansion of phosphodiesterases in the human genome compared with either the worm or fly genomes.

The downstream effectors of the intracellular signaling molecules include the transcription factors that transduce developmental fates. Significant expansions are noted in the ligand-binding nuclear hormone receptor class of transcription factors compared with the fly genome, although not to the extent observed in the worm (Tables 18 and 19). Perhaps the most striking expansion in humans is in the C2H2 zinc finger transcription factors. Pfam detects a total of 4500 C2H2 zinc finger domains in 564 human proteins, compared with 771 in 234 fly proteins. This means that there has been a dramatic expansion not only in the number of C2H2 transcription factors, but also in the number of these DNA-binding motifs per transcription factor (8 on average in humans, 3.3 on average in the fly, and 2.3 on average in the worm). Furthermore, many of these transcription factors contain either the KRAB or SCAN domains, which are not found in the fly or worm genomes. These domains are involved in the oligomerization of transcription factors and increase the combinatorial partnering of these factors. In general, most of the transcription factor domains are shared between the three animal genomes, but the reassortment of these domains results in organism-specific transcription factor families. The domain combinations found in the human, fly, and worm include the BTB with C2H2 in the fly and humans, and homeodomains alone or in combination with Pou and LIM domains in all of the animal genomes. In plants, however, a different set of transcription factors are expanded, namely, the myb family, and a unique set that includes VP1 and AP2 domain–containing proteins (*134*). The yeast genome has a paucity of transcription factors compared with the multicellular eukaryotes, and its repertoire is limited to the expansion of the yeast-specific C6 transcription factor family involved in metabolic regulation.

While we have illustrated expansions in a subset of signal transduction molecules in the human genome compared with the other eukaryotic genomes, it should be noted that most of the protein domains are highly conserved. An interesting observation is that worms and humans have approximately the same number of both tyrosine kinases and serine/threonine kinases (Table 19). It is important to note, however, that these are merely counts of the catalytic domain; the proteins that contain these domains also display a wide repertoire of interaction domains with significant combinatorial diversity.

Hemostasis. Hemostasis is regulated primarily by plasma proteases of the coagulation pathway and by the interactions that occur between the vascular endothelium and platelets. Consistent with known anatomical and physiological differences between vertebrates and invertebrates, extracellular adhesion domains that constitute proteins integral to hemostasis are expanded in the human relative to the fly and worm (Tables 18 and 19). We note the evolution of domains such as FIMAC, FN1, FN2, and C1q that mediate surface interactions between hematopoeitic cells and the vascular matrix. In addition, there has been extensive recruitment of more-ancient animal-specific domains such as VWA, VWC, VWD, kringle, and FN3 into multidomain proteins that are involved in hemostatic regulation. Although we do not find a large expansion in the total number of serine proteases, this enzymatic domain has been specifically recruited into several of these multidomain proteins for proteolytic regulation in the vascular compartment. These are represented in plasma proteins that belong to the kinin and complement pathways. There is a significant expansion in two families of matrix metalloproteases: ADAM (a disintegrin and metalloprotease) and MMPs (matrix metalloproteases) (Table 19). Proteolysis of extracellular matrix (ECM) proteins is critical for tissue development and for tissue degradation in diseases such as cancer, arthritis, Alzheimer's disease, and a variety of inflammatory conditions (*135*, *136*). ADAMs are a family of integral membrane proteins with a pivotal role in fibrinogenolysis and modulating interactions between hematopoietic components and the vascular matrix components. These proteins have been shown to cleave matrix proteins, and even signaling molecules: ADAM-17 converts tumor necrosis factor–α, and ADAM-10 has been implicated in the Notch signaling pathway (*135*). We have identified 19 members of the matrix metalloprotease family, and a total of 51 members of the ADAM and ADAM-TS families.

Apoptosis. Evolutionary conservation of some of the apoptotic pathway components across eukarya is consistent with its central role in developmental regulation and as a response to pathogens and stress signals. The signal transduction pathways involved in programmed cell death, or apoptosis, are mediated by interactions between well-characterized domains that include extracellular domains, adaptor (protein-protein interaction) domains, and those found in effector and regulatory enzymes (*137*). We enumerated the protein counts of central adaptor and effector enzyme domains that are found only in the apoptotic pathways to provide an estimate of divergence across eukarya and relative expansion in the human genome when compared with the fly and worm (Table 18). Adaptor domains found in proteins restricted only to apoptotic regulation such as the DED domains are vertebrate-specific, whereas others like BIR, CARD, and Bcl2 are represented in the fly and worm (although the number of Bcl2 family members in humans is significantly expanded). Although plants and yeast lack the caspases, caspase-like molecules, namely the para- and meta-caspases, have been reported in these organisms (*138*). Compared with other animal genomes, the human genome shows an expansion in the adaptor and effector domain–containing proteins involved in apoptosis, as well as in the proteases involved in the cascade such as the caspase and calpain families.

Expansions of other protein families. *Metabolic enzymes.* There are fewer cytochrome P450 genes in humans than in either the fly or worm. Lipoxygenases (six in humans), on the other hand, appear to be specific to the vertebrates and plants, whereas the lipoxygenase-activating proteins (four in humans) may be vertebrate-specific. Lipoxygenases are involved in arachidonic acid metabolism, and they and their activators have been implicated in diverse human pathology ranging from allergic responses to cancers. One of the most surprising human expansions, however, is in the number of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (46 in humans, 3 in the fly, and 4 in the worm). There is, however, evidence for many retrotransposed GAPDH pseudogenes (*139*), which may account for this apparent expansion. However, it is interesting that GAPDH, long known as a conserved enzyme involved in basic metabolism found across all phyla from bacteria to humans, has recently been shown to have other functions. It has a second cat-

Table 19. Number of proteins assigned to selected Panther families or subfamilies in *H. sapiens* (H), *D. melanogaster* (F), *C. elegans* (W), *S. cerevisiae* (Y), and *A. thaliana* (A).

Panther family/subfamily*	H	F	W	Y	A
Neural structure, function, development					
Ependymin	1	0	0	0	0
Ion channels					
Acetylcholine receptor	17	12	56	0	0
Amiloride-sensitive/degenerin	11	24	27	0	0
CNG/EAG	22	9	9	0	30
IRK	16	3	3	0	0
ITP/ryanodine	10	2	4	0	0
Neurotransmitter-gated	61	51	59	0	19
P2X purinoceptor	10	0	0	0	0
TASK	12	12	48	1	5
Transient receptor	15	3	3	1	0
Voltage-gated Ca^{2+} alpha	22	4	8	2	2
Voltage-gated Ca^{2+} alpha-2	10	3	2	0	0
Voltage-gated Ca^{2+} beta	5	2	2	0	0
Voltage-gated Ca^{2+} gamma	1	0	0	0	0
Voltage-gated K^{+} alpha	33	5	11	0	0
Voltage-gated KQT	6	2	3	0	0
Voltage-gated Na^{+}	11	4	4	9	1
Myelin basic protein	1	0	0	0	0
Myelin PO	5	0	0	0	0
Myelin proteolipid	3	1	0	0	0
Myelin-oligodendrocyte glycoprotein	1	0	0	0	0
Neuropilin	2	0	0	0	0
Plexin	9	2	0	0	0
Semaphorin	22	6	2	0	0
Synaptotagmin	10	3	3	0	0
Immune response					
Defensin	3	0	0	0	0
Cytokine†	86	14	1	0	0
GCSF	1	0	0	0	0
GMCSF	1	0	0	0	0
Intercrine alpha	15	0	0	0	0
Intercrine beta	5	0	0	0	0
Inteferon	8	0	0	0	0
Interleukin	26	1	1	0	0
Leukemia inhibitory factor	1	0	0	0	0
MCSF	1	0	0	0	0
Peptidoglycan recognition protein	2	13	0	0	0
Pre-B cell enhancing factor	1	0	0	0	0
Small inducible cytokine A	14	0	0	0	0
SI cytokine	2	0	0	0	0
TNF	9	0	0	0	0
Cytokine receptor†	62	1	0	0	0
Bradykinin/C-C chemokine receptor	7	0	0	0	0
Fl cytokine receptor	2	0	0	0	0
Interferon receptor	3	0	0	0	0
Interleukin receptor	32	0	0	0	0
Leukocyte tyrosine kinase receptor	3	0	0	0	0
MCSF receptor	1	0	0	0	0
TNF receptor	3	0	0	0	0
Immunoglobulin receptor†	59	0	0	0	0
T-cell receptor alpha chain	16	0	0	0	0
T-cell receptor beta chain	15	0	0	0	0
T-cell receptor gamma chain	1	0	0	0	0
T-cell receptor delta chain	1	0	0	0	0
Immunoglobulin FC receptor	8	0	0	0	0
Killer cell receptor	16	0	0	0	0
Polymeric-immunoglobulin receptor	4	0	0	0	0

alytic activity, as a uracil DNA glycosylase (*140*) and functions as a cell cycle regulator (*141*) and has even been implicated in apoptosis (*142*).

Translation. Another striking set of human expansions has occurred in certain families involved in the translational machinery. We identified 28 different ribosomal subunits that each have at least 10 copies in the genome; on average, for all ribosomal proteins there is about an 8- to 10-fold expansion in the number of genes relative to either the worm or fly. Retrotransposed pseudogenes may account for many of these expansions [see the discussion above and (*143*)]. Recent evidence suggests that a number of ribosomal proteins have secondary functions independent of their involvement in protein biosynthesis; for example, L13a and the related L7 subunits (36 copies in humans) have been shown to induce apoptosis (*144*).

There is also a four- to fivefold expansion in the elongation factor 1-alpha family (eEF1A; 56 human genes). Many of these expansions likely represent intronless paralogs that have presumably arisen from retrotransposition, and again there is evidence that many of these may be pseudogenes (*145*). However, a second form (eEF1A2) of this factor has been identied with tissue-specific expression in skeletal muscle and a complementary expression pattern to the ubiquitously expressed eEF1A (*146*).

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
MHC class I	22	0	0	0	0
MHC class II	20	0	0	0	0
Other immunoglobulin†	114	0	0	0	0
Toll receptor–related	10	6	0	0	0
Developmental and homeostatic regulators					
Signaling molecules†					
Calcitonin	3	0	0	0	0
Ephrin	8	2	4	0	0
FGF	24	1	1	0	0
Glucagon	4	0	0	0	0
Glycoprotein hormone beta chain	2	0	0	0	0
Insulin	1	0	0	0	0
Insulin-like hormone	3	0	0	0	0
Nerve growth factor	3	0	0	0	0
Neuregulin/heregulin	6	0	0	0	0
neuropeptide Y	4	0	0	0	0
PDGF	1	1	0	0	0
Relaxin	3	0	0	0	0
Stannocalcin	2	0	0	0	0
Thymopoeitin	2	0	1	0	0
Thyomosin beta	4	2	0	0	0
TGF-β	29	6	4	0	0
VEGF	4	0	0	0	0
Wnt	18	6	5	0	0
Receptors†					
Ephrin receptor	12	2	1	0	0
FGF receptor	4	4	0	0	0
Frizzled receptor	12	6	5	0	0
Parathyroid hormone receptor	2	0	0	0	0
VEGF receptor	5	0	0	0	0
BDNF/NT-3 nerve growth factor receptor	4	0	0	0	0
Kinases and phosphatases					
Dual-specificity protein phosphatase	29	8	10	4	11
S/T and dual-specificity protein kinase†	395	198	315	114	1102
S/T protein phosphatase	15	19	51	13	29
Y protein kinase†	106	47	100	5	16
Y protein phosphatase	56	22	95	5	6
Signal transduction					
ARF family	55	29	27	12	45
Cyclic nucleotide phosphodiesterase	25	8	6	1	0
G protein-coupled receptors†‡	616	146	284	0	1
G-protein alpha	27	10	22	2	5
G-protein beta	5	3	2	1	1
G-protein gamma	13	2	2	0	0
Ras superfamily	141	64	62	26	86
G-protein modulators†					
ARF GTPase-activating	20	8	9	5	15
Neurofibromin	7	2	0	2	0
Ras GTPase-activating	9	3	8	1	0
Tuberin	7	3	2	0	0
Vav proto-oncogene family	35	15	13	3	0

Ribonucleoproteins. Alternative splicing results in multiple transcripts from a single gene, and can therefore generate additional diversity in an organism's protein complement. We have identified 269 genes for ribonucleoproteins. This represents over 2.5 times the number of ribonucleoprotein genes in the worm, two times that of the fly, and about the same as the 265 identified in the *Arabidopsis* genome. Whether the diversity of ribonucleoprotein genes in humans contributes to gene regulation at either the splicing or translational level is unknown.

Posttranslational modifications. In this set of processes, the most prominent expansion is the transglutaminases, calcium-dependent enzymes that catalyze the cross-linking of proteins in cellular processes such as hemostasis and apoptosis (*147*). The vitamin K–dependent gamma carboxylase gene product acts on the GLA domain (missing in the fly and worm) found in coagulation factors, osteocalcin, and matrix GLA protein (*148*). Tyrosylprotein sulfotransferases participate in the posttranslational modification of proteins involved in inflammation and hemostasis, including coagulation factors and chemokine receptors (*149*). Although there is no significant numerical increase in the counts for domains involved in nuclear protein modification, there are a number of domain arrangements in the predicted human proteins that are not found in the other currently sequenced genomes. These include the tandem association of two histone deacetylase domains in HD6 with a ubiquitin finger domain, a feature lacking in the fly genome. An additional example is the co-occurrence of important nuclear regulatory enzyme PARP (poly-ADP ribosyl transferase) domain fused to protein-interaction domains—BRCT and VWA in humans.

Concluding remarks. There are several possible explanations for the differences in phenotypic complexity observed in humans when compared to the fly and worm. Some of these relate to the prominent differences in the immune system, hemostasis, neuronal, vascular, and cytoskeletal complexity. The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the levels of protein architecture, transcriptional and translational control, posttranslational modification of proteins, or posttranscriptional regulation. Extensive domain shuffling to increase or alter combinatorial diversity can provide an exponential increase in the ability to mediate protein-protein interactions without dramatically increasing the absolute size of the protein complement (*150*). Evolution of apparently new (from the perspective of sequence analysis) protein domains and increasing regulatory complexity by domain accretion both quantitatively and qualitatively (recruitment of novel domains with preexisting ones) are two features that we observe in humans. Perhaps the best illustration of this trend is the C2H2 zinc finger–containing transcription factors, where we see expansion in the number of domains per protein, together with vertebrate-specific domains such as KRAB and SCAN. Recent reports on the prominent use of internal ribosomal entry sites in the human genome to regulate translation of specific classes of proteins suggests that this is an area that needs further research to identify the full extent of this process in the human genome (*151*). At the posttranslational level, although we provide examples of expansions of some protein families involved in these modifications, further experimental evidence is required to evaluate whether this is correlated with increased complexity in protein processing. Posttranscriptional processing and the extent of isoform generation in the human remain to be cataloged in their entirety. Given the conserved nature of the spliceosomal machinery, further analysis will be required to dissect regulation at this level.

8 Conclusions

8.1 The whole-genome sequencing approach versus BAC by BAC

Experience in applying the whole-genome shotgun sequencing approach to a diverse group of organisms with a wide range of genome sizes and repeat content allows us to assess its strengths and weaknesses. With the success of the method for a large number of microbial genomes, *Drosophila*, and now the human, there can be no doubt concerning the utility of this method. The large number of microbial genomes that have been sequenced by this method (*15, 80, 152*) demonstrate that megabase-sized genomes can be sequenced efficiently without any input other that the de novo mate-paired sequences. With more complex genomes like those of *Drosophila* or human, map information, in the form of well-ordered markers, has been critical for long-range ordering of scaffolds. For joining scaffolds into chromosomes, the quality of the map (in terms of the order of the markers) is more important than the number of markers per se. Although this mapping could have been performed concurrently with sequencing, the prior existence of mapping data was beneficial. During the sequencing of the *A. thaliana* genome, sequencing of individual BAC clones permitted extension of the se-

Table 19 (*Continued*)

Panther family/subfamily*	H	F	W	Y	A
	Transcription factors/chromatin organization				
C2H2 zinc finger–containing†	607	232	79	28	8
COE	7	1	1	0	0
CREB	7	1	2	0	0
ETS-related	25	8	10	0	0
Forkhead-related	34	19	15	4	0
FOS	8	2	1	0	0
Groucho	13	2	1	0	0
Histone H1	5	0	1	0	0
Histone H2A	24	1	17	3	13
Histone H2B	21	1	17	2	12
Histone H3	28	2	24	2	16
Histone H4	9	1	16	1	8
Homeotic†	168	104	74	4	78
ABD-B	5	0	0	0	0
Bithoraxoid	1	8	1	0	0
Iroquois class	7	3	1	0	0
Distal-less	5	2	1	0	0
Engrailed	2	2	1	0	0
LIM-containing	17	8	3	0	0
MEIS/KNOX class	9	4	4	2	26
NK-3/NK-2 class	9	4	5	0	0
Paired box	38	28	23	0	2
Six	5	3	4	0	0
Leucine zipper	6	0	0	0	0
Nuclear hormone receptor†	59	25	183	1	4
Pou-related	15	5	4	1	0
Runt-related	3	4	2	0	0
	ECM adhesion				
Cadherin	113	17	16	0	0
Claudin	20	0	0	0	0
Complement receptor-related	22	8	6	0	0
Connexin	14	0	0	0	0
Galectin	12	5	22	0	0
Glypican	13	2	1	0	0
ICAM	6	0	0	0	0
Integrin alpha	24	7	4	0	1
Integrin beta	9	2	2	0	0
LDL receptor family	26	19	20	0	2
Proteoglycans	22	9	7	0	5
	Apoptosis				
Bcl-2	12	1	0	0	0
Calpain	22	4	11	1	3
Calpain inhibitor	4	0	0	0	1
Caspase	13	7	3	0	0
	Hemostasis				
ADAM/ADAMTS	51	9	12	0	0
Fibronectin	3	0	0	0	0
Globin	10	2	3	0	3
Matrix metalloprotease	19	2	7	0	3
Serum amyloid A	4	0	0	0	0
Serum amyloid P (subfamily of Pentaxin)	2	0	0	0	0
Serum paraoxonase/arylesterase	4	0	3	0	0
Serum albumin	4	0	0	0	0
Transglutaminase	10	1	0	0	0
	Other enzymes				
Cytochrome p450	60	89	83	3	256
GAPDH	46	3	4	3	8
Heparan sulfotransferase	11	4	2	0	0
	Splicing and translation				
EF-1alpha	56	13	10	6	13
Ribonucleoproteins†	269	135	104	60	265
Ribosomal proteins†	812	111	80	117	256

*The table lists Panther families or subfamilies relevant to the text that either (i) are not specifically represented by Pfam (Table 18) or (ii) differ in counts from the corresponding Pfam models. †This class represents a number of different families in the same Panther molecular function subcategory. ‡This count includes only rhodopsin-class, secretin-class, and metabotropic glutamate-class GPCRs.

quence well into centromeric regions and allowed high-quality resolution of complex repeat regions. Likewise, in *Drosophila*, the BAC physical map was most useful in regions near the highly repetitive centromeres and telomeres. WGA has been found to deliver excellent-quality reconstructions of the unique regions of the genome. As the genome size, and more importantly the repetitive content, increases, the WGA approach delivers less of the repetitive sequence.

The cost and overall efficiency of clone-by-clone approaches makes them difficult to justify as a stand-alone strategy for future large-scale genome-sequencing projects. Specific applications of BAC-based or other clone mapping and sequencing strategies to resolve ambiguities in sequence assembly that cannot be efficiently resolved with computational approaches alone are clearly worth exploring. Hybrid approaches to whole-genome sequencing will only work if there is sufficient coverage in both the whole-genome shotgun phase and the BAC clone sequencing phase. Our experience with human genome assembly suggests that this will require at least 3× coverage of both whole-genome and BAC shotgun sequence data.

8.2 The low gene number in humans

We have sequenced and assembled ~95% of the euchromatic sequence of *H. sapiens* and used a new automated gene prediction method to produce a preliminary catalog of the human genes. This has provided a major surprise: We have found far fewer genes (26,000 to 38,000) than the earlier molecular predictions (50,000 to over 140,000). Whatever the reasons for this current disparity, only detailed annotation, comparative genomics (particularly using the *Mus musculus* genome), and careful molecular dissection of complex phenotypes will clarify this critical issue of the basic "parts list" of our genome. Certainly, the analysis is still incomplete and considerable refinement will occur in the years to come as the precise structure of each transcription unit is evaluated. A good place to start is to determine why the gene estimates derived from EST data are so discordant with our predictions. It is likely that the following contribute to an inflated gene number derived from ESTs: the variable lengths of 3′- and 5′-untranslated leaders and trailers; the little-understood vagaries of RNA processing that often leave intronic regions in an unspliced condition; the finding that nearly 40% of human genes are alternatively spliced (*153*); and finally, the unsolved technical problems in EST library construction where contamination from heterogeneous nuclear RNA and genomic DNA are not uncommon. Of course, it is possible that there are genes that remain unpredicted owing to the absence of EST or protein data to support them, although our use of mouse genome data for predicting genes should limit this number. As was true at the beginning of genome sequencing, ultimately it will be necessary to measure mRNA in specific cell types to demonstrate the presence of a gene.

J. B. S. Haldane speculated in 1937 that a population of organisms might have to pay a price for the number of genes it can possibly carry. He theorized that when the number of genes becomes too large, each zygote carries so many new deleterious mutations that the population simply cannot maintain itself. On the basis of this premise, and on the basis of available mutation rates and x-ray–induced mutations at specific loci, Muller, in 1967 (*154*), calculated that the mammalian genome would contain a maximum of not much more than 30,000 genes (*155*). An estimate of 30,000 gene loci for humans was also arrived at by Crow and Kimura (*156*). Muller's estimate for *D. melanogaster* was 10,000 genes, compared to 13,000 derived by annotation of the fly genome (*26, 27*). These arguments for the theoretical maximum gene number were based on simplified ideas of genetic load—that all genes have a certain low rate of mutation to a deleterious state. However, it is clear that many mouse, fly, worm, and yeast knockout mutations lead to almost no discernible phenotypic perturbations.

The modest number of human genes means that we must look elsewhere for the mechanisms that generate the complexities inherent in human development and the sophisticated signaling systems that maintain homeostasis. There are a large number of ways in which the functions of individual genes and gene products are regulated. The degree of "openness" of chromatin structure and hence transcriptional activity is regulated by protein complexes that involve histone and DNA enzymatic modifications. We enumerate many of the proteins that are likely involved in nuclear regulation in Table 19. The location, timing, and quantity of transcription are intimately linked to nuclear signal transduction events as well as by the tissue-specific expression of many of these proteins. Equally important are regulatory DNA elements that include insulators, repeats, and endogenous viruses (*157*); methylation of CpG islands in imprinting (*158*); and promoter-enhancer and intronic regions that modulate transcription. The spliceosomal machinery consists of multisubunit proteins (Table 19) as well as structural and catalytic RNA elements (*159*) that regulate transcript structure through alternative start and termination sites and splicing. Hence, there is a need to study different classes of RNA molecules (*160*) such as small nucleolar RNAs, antisense riboregulator RNA, RNA involved in X-dosage compensation, and other structural RNAs to appreciate their precise role in regulating gene expression. The phenomenon of RNA editing in which coding changes occur directly at the level of mRNA is of clinical and biological relevance (*161*). Finally, examples of translational control include internal ribosomal entry sites that are found in proteins involved in cell cycle regulation and apoptosis (*162*). At the protein level, minor alterations in the nature of protein-protein interactions, protein modifications, and localization can have dramatic effects on cellular physiology (*163*). This dynamic system therefore has many ways to modulate activity, which suggests that definition of complex systems by analysis of single genes is unlikely to be entirely successful.

In situ studies have shown that the human genome is asymmetrically populated with G+C content, CpG islands, and genes (*68*). However, the genes are not distributed quite as unequally as had been predicted (Table 9) (*69*). The most G+C-rich fraction of the genome, H3 isochores, constitute more of the genome than previously thought (about 9%), and are the most gene-dense fraction, but contain only 25% of the genes, rather than the predicted ~40%. The low G+C L isochores make up 65% of the genome, and 48% of the genes. This inhomogeneity, the net result of millions of years of mammalian gene duplication, has been described as the "desertification" of the vertebrate genome (*71*). Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution? If these deserts are dispensable, it ought to be possible to find mammalian genomes that are far smaller in size than the human genome. Indeed, many species of bats have genome sizes that are much smaller than that of humans; for example, *Miniopterus*, a species of Italian bat, has a genome size that is only 50% that of humans (*164*). Similarly, *Muntiacus*, a species of Asian barking deer, has a genome size that is ~70% that of humans.

8.3 Human DNA sequence variation and its distribution across the genome

This is the first eukaryotic genome in which a nearly uniform ascertainment of polymorphism has been completed. Although we have identified and mapped more than 3 million SNPs, this by no means implies that the task of finding and cataloging SNPs is complete. These represent only a fraction of the SNPs present in the human population as a whole. Nevertheless, this first glimpse at genome-wide variation has revealed strong inhomogeneities in the distribution of SNPs across the genome. Polymorphism in DNA carries with it a snapshot of the past operation of population genetic forces, including mutation, migration, selection, and genetic drift. The availability of a dense array of SNPs will allow questions related to each of these factors to be addressed on a genome-wide basis. SNP studies can establish the range of haplotypes present in subjects of different ethnogeographic origins, providing insights into population history and migration patterns. Although such studies have suggested that modern human lineages derive from Africa, many important questions regarding human origins remain unanswered, and more analyses using detailed SNP maps will be needed to settle these controversies. In addition to providing evidence for population expansions, migration, and admixture, SNPs can serve as markers for the extent of evolutionary constraint acting on particular genes. The correlation between patterns of intraspecies and interspecies genetic variation may prove to be especially informative to identify sites of reduced genetic diversity that may mark loci where sequence variations are not tolerated.

The remarkable heterogeneity in SNP density implies that there are a variety of forces acting on polymorphism—sparse regions may have lower SNP density because the mutation rate is lower, because most of those regions have a lower fraction of mutations that are tolerated, or because recent strong selection in favor of a newly arisen allele "swept" the linked variation out of the population (*165*). The effect of random genetic drift also varies widely across the genome. The nonrecombining portion of the Y chromosome faces the strongest pressure from random drift because there are roughly one-quarter as many Y chromosomes in the population as there are autosomal chromosomes, and the level of polymorphism on the Y is correspondingly less. Similarly, the X chromosome has a smaller effective population size than the autosomes, and its nucleotide diversity is also reduced. But even across a single autosome, the effective population size can vary because the density of deleterious mutations may vary. Regions of high density of deleterious mutations will see a greater rate of elimination by selection, and the effective population size will be smaller (*166*). As a result, the density of even completely neutral SNPs will be lower in such regions. There is a large literature on the association between SNP density and local recombination rates in *Drosophila*, and it remains an important task to assess the strength of this association in the human genome, because of its impact on the design of local SNP densities for disease-association studies. It also remains an important task to validate SNPs on a genomic scale in order to assess the degree of heterogeneity among geographic and ethnic populations.

8.4 Genome complexity

We will soon be in a position to move away from the cataloging of individual components of the system, and beyond the simplistic notions of "this binds to that, which then docks on this, and then the complex moves there. . . ." (*167*) to the exciting area of network perturbations, nonlinear responses and thresholds, and their pivotal role in human diseases.

The enumeration of other "parts lists" reveals that in organisms with complex nervous systems, neither gene number, neuron number, nor number of cell types correlates in any meaningful manner with even simplistic measures of structural or behavioral complexity. Nor would they be expected to; this is the realm of nonlinearities and epigenesis (*168*). The 520 million neurons of the common octopus exceed the neuronal number in the brain of a mouse by an order of magnitude. It is apparent from a comparison of genomic data on the mouse and human, and from comparative mammalian neuroanatomy (*169*), that the morphological and behavioral diversity found in mammals is underpinned by a similar gene repertoire and similar neuroanatomies. For example, when one compares a pygmy marmoset (which is only 4 inches tall and weighs about 6 ounces) to a chimpanzee, the brain volume of this minute primate is found to be only about 1.5 cm^3, two orders of magnitude less than that of a chimp and three orders less than that of humans. Yet the neuroanatomies of all three brains are strikingly similar, and the behavioral characteristics of the pygmy marmoset are little different from those of chimpanzees. Between humans and chimpanzees, the gene number, gene structures and functions, chromosomal and genomic organizations, and cell types and neuroanatomies are almost indistinguishable, yet the developmental modifications that predisposed human lineages to cortical expansion and development of the larynx, giving rise to language, culminated in a massive singularity that by even the simplest of criteria made humans more complex in a behavioral sense.

Simple examination of the number of neurons, cell types, or genes or of the genome size does not alone account for the differences in complexity that we observe. Rather, it is the interactions within and among these sets that result in such great variation. In addition, it is possible that there are "special cases" of regulatory gene networks that have a disproportionate effect on the overall system. We have presented several examples of "regulatory genes" that are significantly increased in the human genome compared with the fly and worm. These include extracellular ligands and their cognate receptors (e.g., wnt, frizzled, TGF-β, ephrins, and connexins), as well as nuclear regulators (e.g., the KRAB and homeodomain transcription factor families), where a few proteins control broad developmental processes. The answers to these "complexities" perhaps lie in these expanded gene families and differences in the regulatory control of ancient genes, proteins, pathways, and cells.

8.5 Beyond single components

While few would disagree with the intuitive conclusion that Einstein's brain was more complex than that of *Drosophila*, closer comparisons such as whether the set of predicted human proteins is more complex than the protein set of *Drosophila*, and if so, to what degree, are not straightforward, since protein, protein domain, or protein-protein interaction measures do not capture context-dependent interactions that underpin the dynamics underlying phenotype.

Currently, there are more than 30 different mathematical descriptions of complexity (*170*). However, we have yet to understand the mathematical dependency relating the number of genes with organism complexity. One pragmatic approach to the analysis of biological systems, which are composed of nonidentical elements (proteins, protein complexes, interacting cell types, and interacting neuronal populations), is through graph theory (*171*). The elements of the system can be represented by the vertices of complex topographies, with the edges representing the interactions between them. Examination of large networks reveals that they can self-organize, but more important, they can be particularly robust. This robustness is not due to redundancy, but is a property of inhomogeneously wired networks. The error tolerance of such networks comes with a price; they are vulnerable to the selection or removal of a few nodes that contribute disproportionately to network stability. Gene knockouts provide an illustration. Some knockouts may have minor effects, whereas others have catastrophic effects on the system. In the case of vimentin, a supposedly critical component of the cytoplasmic intermediate filament network of mammals, the knockout of the gene in mice reveals them to be reproductively normal, with no obvious phenotypic effects (*172*), and yet the usually conspicuous vimentin network is completely absent. On the other hand, ~30% of knockouts in *Drosophila* and mice correspond to critical nodes whose reduction in gene product, or total elimination, causes the network to crash most of the time, although even in some of these cases, phenotypic normalcy ensues, given the appropriate genetic background. Thus, there are no "good" genes or "bad" genes, but only networks that exist at various levels and at different connectivities, and at different states of sensitivity to perturbation. Sophisticated mathematical analysis needs to be constantly evaluated against hard biological data sets that specifically address network dynamics. Nowhere is this more critical than in attempts to come to grips with "complexity," particularly because deconvoluting and correcting complex networks that have undergone perturbation, and have resulted in human diseases, is the greatest significant challenge now facing us.

It has been predicted for the last 15 years that complete sequencing of the human genome would open up new strategies for human biological research and would have a major impact on medicine, and through medicine and public health, on society. Effects on biomedical research are already being felt. This assembly of the human genome sequence is but a first, hesitant step on a long and exciting journey toward understanding the role of the genome in human biology. It has been possible only because of innovations in instrumentation and software that have allowed automation of almost every step of the process from DNA preparation to annotation. The next steps are clear: We must define the complexity that ensues when this relatively modest set of about 30,000 genes is expressed. The sequence provides the framework upon which all the genetics, biochemistry, physiology, and ultimately phenotype depend. It provides the boundaries for scientific inquiry. The sequence is only the first level of understanding of the genome. All genes and their control elements must be identified; their functions, in concert as well as in isolation, defined; their sequence variation worldwide described; and the relation between genome variation and specific phenotypic characteristics determined. Now we know what we have to explain.

Another paramount challenge awaits: public discussion of this information and its potential for improvement of personal health. Many diverse sources of data have shown that any two individuals are more than 99.9% identical in sequence, which means that all the glorious differences among individuals in our species that can be attributed to genes falls in a mere 0.1% of the sequence. There are two fallacies to be avoided: determinism, the idea that all characteristics of the person are "hard-wired" by the genome; and reductionism, the view that with complete knowledge of the human genome sequence, it is only a matter of time before our understanding of gene functions and interactions will provide a complete causal description of human variability. The real challenge of human biology, beyond the task of finding out how genes orchestrate the construction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.

References and Notes

1. R. L. Sinsheimer, *Genomics* **5**, 954 (1989); U.S. Department of Energy, Office of Health and Environmental Research, *Sequencing the Human Genome: Summary Report of the Santa Fe Workshop*, Santa Fe, NM, 3 to 4 March 1986 (Los Alamos National Laboratory, Los Alamos, NM, 1986).
2. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1996).
3. F. Sanger *et al.*, *Nature* **265**, 687 (1977).
4. P. H. Seeburg *et al.*, *Trans. Assoc. Am. Physicians* **90**, 109 (1977).
5. E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, *Anal. Biochem.* **154**, 353 (1986).
6. J. Gocayne *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8296 (1987).
7. A. Martin-Gallardo *et al.*, *DNA Sequence* **3**, 237 (1992); W. R. McCombie *et al.*, *Nature Genet.* **1**, 348 (1992); M. A. Jensen *et al.*, *DNA Sequence* **1**, 233 (1991).
8. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
9. M. D. Adams *et al.*, *Nature* **355**, 632 (1992); M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 256 (1993); M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, *Nature Genet.* **4**, 373 (1993); M. H. Polymeropoulos *et al.*, *Nature Genet.* **4**, 381 (1993); M. Marra *et al.*, *Nature Genet.* **21**, 191 (1999).
10. M. D. Adams *et al.*, *Nature* **377**, 3 (1995); O. White *et al.*, *Nucleic Acids Res.* **21**, 3829 (1993).
11. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 729 (1982).
12. B. W. J. Mahy, J. J. Esposito, J. C. Venter, *Am. Soc. Microbiol. News* **57**, 577 (1991).
13. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
14. C. M. Fraser *et al.*, *Science* **270**, 397 (1995).
15. C. J. Bult *et al.*, *Science* **273**, 1058 (1996); J. F. Tomb *et al.*, *Nature* **388**, 539 (1997); H. P. Klenk *et al.*, *Nature* **390**, 364 (1997).
16. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
17. H. Schmitt *et al.*, *Genomics* **33**, 9 (1996).
18. S. Zhao *et al.*, *Genomics* **63**, 321 (2000).
19. X. Lin *et al.*, *Nature* **402**, 761 (1999).
20. J. L. Weber, E. W. Myers, *Genome Res.* **7**, 401 (1997).
21. P. Green, *Genome Res.* **7**, 410 (1997).
22. E. Pennisi, *Science* **280**, 1185 (1998).
23. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
24. M. D. Adams *et al.*, *Nature* **368**, 474 (1994).
25. E. Marshall, E. Pennisi, *Science* **280**, 994 (1998).
26. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
27. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
28. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
29. F. S. Collins *et al.*, *Science* **282**, 682 (1998).
30. International Human Genome Sequencing Consortium (2001), *Nature* **409**, 860 (2001).
31. Institutional review board: P. Calabresi (chairman), H. P. Freeman, C. McCarthy, A. L. Caplan, G. D. Rogell, J. Karp, M. K. Evans, B. Margus, C. L. Carter, R. A. Millman, S. Broder.
32. Eligibility criteria for participation in the study were as follows: prospective donors had to be 21 years of age or older, not pregnant, and capable of giving an informed consent. Donors were asked to self-define their ethnic backgrounds. Standard blood bank screens (screening for HIV, hepatitis viruses, and so forth) were performed on all samples at the clinical laboratory prior to DNA extraction in the Celera laboratory. All samples that tested positive for transmissible viruses were ineligible and were discarded. Karyotype analysis was performed on peripheral blood lymphocytes from all samples selected for sequencing; all were normal. A two-staged consent process for prospective donors was employed. The first stage of the consent process provided information about the genome project, procedures, and risks and benefits of participating. The second stage of the consent process involved answering follow-up questions and signing consent forms, and was conducted about 48 hours after the first.
33. DNA was isolated from blood (*173*) or sperm. For sperm, a washed pellet (100 μl) was lysed in a suspension (1 ml) containing 0.1 M NaCl, 10 mM tris-Cl–20 mM EDTA (pH 8), 1% SDS, 1 mg proteinase K, and 10 mM dithiothreitol for 1 hour at 37°C. The lysate was extracted with aqueous phenol and with phenol/chloroform. The DNA was ethanol precipitated and dissolved in 1 ml TE buffer. To make genomic libraries, DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3′-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3′-TGTG overhangs. Libraries with three different average sizes of inserts were constructed: 2, 10, and 50 kbp. The 2-kbp fragments were cloned in a high-copy pUC18 derivative. The 10- and 50-kbp fragments were cloned in a medium-copy pBR322 derivative. The 2- and 10-kbp libraries yielded uniform-sized large colonies on plating. However, the 50-kbp libraries produced many small colonies and inserts were unstable. To remedy this, the 50-kbp libraries were digested with Bgl II, which does not cleave the vector, but generally cleaved several times within the 50-kbp insert. A 1264-bp Bam HI kanamycin resistance cassette (purified from pUCK4; Amersham Pharmacia, catalog no. 27-4958-01) was added and ligation was carried out at 37°C in the continual presence of Bgl II. As Bgl II–Bgl II ligations occurred, they were continually cleaved, whereas Bam HI–Bgl II ligations were not cleaved. A high yield of internally deleted circular library molecules was obtained in which the residual insert ends were separated by the kanamycin cassette DNA. The internally deleted libraries, when plated on agar containing ampicillin (50 μg/ml), carbenicillin (50 μg/ml), and kanamycin (15 μg/ml), produced relatively uniform large colonies. The resulting clones could be prepared for sequencing using the same procedures as clones from the 10-kbp libraries.
34. Transformed cells were plated on agar diffusion plates prepared with a fresh top layer containing no antibiotic poured on top of a previously set bottom layer containing excess antibiotic, to achieve the correct final concentration. This method of plating permitted the cells to develop antibiotic resistance before being exposed to antibiotic without the potential clone bias that can be introduced through liquid outgrowth protocols. After colonies had grown, QBot (Genetix, UK) automated colony-picking robots were used to pick colonies meeting stringent size and shape criteria and to inoculate 384-well microtiter plates containing liquid growth medium. Liquid cultures were incubated overnight, with shaking, and were scored for growth before passing to template preparation. Template DNA was extracted from liquid bacterial culture using a procedure based upon the alkaline lysis miniprep method (*173*) adapted for high throughput processing in 384-well microtiter plates. Bacterial cells were lysed; cell debris was removed by centrifugation; and plasmid DNA was recovered by isopropanol precipitation and resuspended in 10 mM tris-HCl buffer. Reagent dispensing operations were accomplished using Titertek MAP 8 liquid dispensing systems. Plate-to-plate liquid transfers were performed using Tomtec Quadra 384 Model 320 pipetting robots. All plates were tracked throughout processing by unique plate barcodes. Mated sequencing reads from opposite ends of each clone insert were obtained by preparing two 384-well cycle sequencing reaction plates from each plate of plasmid template DNA using ABI-PRISM BigDye Terminator chemistry (Applied Biosystems) and standard M13 forward and reverse primers. Sequencing reactions were prepared using the Tomtec Quadra 384-320 pipetting robot. Parent-child plate relationships and, by extension, forward-reverse sequence mate pairs were established by automated plate barcode reading by the onboard barcode reader and were recorded by direct LIMS communication. Sequencing reaction products were purified by alcohol precipitation and were dried, sealed, and stored at 4°C in the dark until needed for sequencing, at which time the reaction products were resuspended in deionized formamide and sealed immediately to prevent degradation. All sequence data were generated using a single sequencing platform, the ABI PRISM 3700 DNA Analyzer. Sample sheets were created at load time using a Java-based application that facilitates barcode scanning of the sequencing plate barcode, retrieves sample information from the central LIMS, and reserves unique trace identifiers. The application permitted a single sample sheet file in the linking directory and deleted previously created sample sheet files immediately upon scanning of a

sample plate barcode, thus enhancing sample s. to-plate associations.

35. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977); J. M. Prober *et al.*, *Science* **238**, 336 (1987).
36. Celera's computing environment is based on Compaq Computer Corporation's Alpha system technology running the Tru64 Unix operating system. Celera uses these Alphas as Data Servers and as nodes in a Virtual Compute Farm, all of which are connected to a fully switched network operating at Fast Ethernet speed (for the VCF) and gigabit Ethernet speed (for data servers). Load balancing and scheduling software manages the submission and execution of jobs, based on central processing unit (CPU) speed, memory requirements, and priority. The Virtual Compute Farm is composed of 440 Alpha CPUs, which includes model EV6 running at a clock speed of 400 MHz and EV67 running at 667 MHz. Available memory on these systems ranges from 2 GB to 8 GB. The VCF is used to manage trace file processing, and annotation. Genome assembly was performed on a GS 160 running 16 EV67s (667 MHz) and 64 GB of memory, and 10 ES40s running 4 EV6s (500 MHz) and 32 GB of memory. A total of 100 terabytes of physical disk storage was included in a Storage Area Network that was available to systems across the environment. To ensure high availability, file and database servers were configured as 4-node Alpha TruClusters, so that services would fail over in the event of hardware or software failure. Data availability was further enhanced by using hardware- and software-based disk mirroring (RAID-0), disk striping (RAID-1), and disk striping with parity (RAID-5).
37. Trace processing generates quality values for base calls by means of Paracel's TraceTuner, trims sequence reads according to quality values, trims vector and adapter sequence from high-quality reads, and screens sequences for contaminants. Similar in design and algorithm to the phred program (*174*), TraceTuner reports quality values that reflect the log-odds score of each base being correct. Read quality was evaluated in 50-bp windows, each read being trimmed to include only those consecutive 50-bp segments with a minimum mean accuracy of 97%. End windows (both ends of the trace) of 1, 5, 10, 25, and 50 bases were trimmed to a minimum mean accuracy of 98%. Every read was further checked for vector and contaminant matches of 50 bp or more, and if found, the read was removed from consideration. Finally, any match to the 5′ vector splice junction in the initial part of a read was removed.
38. National Center for Biotechnology Information (NCBI); available at www.ncbi.nlm.nih.gov/.
39. NCBI; available at www.ncbi.nlm.nih.gov/HTGS/.
40. All bactigs over 3 kbp were examined for coverage by Celera mate pairs. An interval of a bactig was deemed an assembly error where there were no mate pairs spanning the interval and at least two reads that should have their mate on the other side of the interval but did not. In other words, there was no mate pair evidence supporting a join in the breakpoint interval and at least two mate pairs contradicting the join. By this criterion, we detected and broke apart bactigs at 13,037 locations, or equivalently, we found 2.13% of the bactigs to be misassembled.
41. We considered a BAC entry to be chimeric if, by the Lander-Waterman statistic (*175*), the odds were 0.99 or more that the assembly we produced was inconsistent with the sequence coming from a single source. By this criterion, 714 or 2.2% of BAC entries were deemed chimeric.
42. G. Myers, S. Selznick, Z. Zhang, W. Miller, *J. Comput. Biol.* **3**, 563 (1996).
43. E. W. Myers, J. L. Weber, in *Computational Methods in Genome Research*, S. Suhai, Ed. (Plenum, New York, 1996), pp. 73–89.
44. P. Deloukas *et al.*, *Science* **282**, 744 (1998).
45. M. A. Marra *et al.*, *Genome Res.* **7**, 1072 (1997).
46. J. Zhang *et al.*, data not shown.
47. Shredded bactigs were located on long CSA scaffolds (>500 kbp) and the distribution of these fragments on the scaffolds was analyzed. If the spread of these fragments was greater than four times the reported BAC length, the BAC was considered to be chimeric. In addition, if >20% of bactigs of a given BAC were found on a different scaffolds that were not adjacent in map position, then the BAC was also considered as chimeric. The total chimeric BACs divided by the number of BACs used for CSA gave the minimal estimate of chimerism rate.
48. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
49. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
50. A. B. Carvalho, B. P. Lazzaro, A. G. Clark, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13239 (2000).
51. The International RH Mapping Consortium, available at www.ncbi.nlm.nih.gov/genemap99/.
52. See http://ftp.genome.washington.edu/RM/RepeatMasker.html.
53. G. D. Schuler, *Trends Biotechnol.* **16**, 456 (1998).
54. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).

55a. M. Olivier *et al.*, *Science* **291**, 1298 (2001).

55b. See http://genome.ucsc.edu/.

56. N. Chaudhari, W. E. Hahn, *Science* **220**, 924 (1983); R. J. Milner, J. G. Sutcliffe, *Nucleic Acids Res.* **11**, 5497 (1983).
57. D. Dickson, *Nature* **401**, 311 (1999).
58. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
59. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
60. M. Yandell, in preparation.
61. K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, *Trends Genet.* **16**, 44 (2000).
62. Scaffolds containing greater than 10 kbp of sequence were analyzed for features of biological importance through a series of computational steps, and the results were stored in a relational database. For scaffolds greater than one megabase, the sequence was cut into single megabase pieces before computational analysis. All sequence was masked for complex repeats using Repeatmasker (*52*) before gene finding or homology-based analysis. The computational pipeline required ~7 hours of CPU time per megabase, including repeat masking, or a total compute time of about 20,000 CPU hours. Protein searches were performed against the nonredundant protein database available at the NCBI. Nucleotide searches were performed against human, mouse, and rat Celera Gene Indices (assemblies of cDNA and EST sequences), mouse genomic DNA reads generated at Celera (3×), the Ensembl gene database available at the European Bioinformatics Institute (EBI), human and rodent (mouse and rat) EST data sets parsed from the dbEST database (NCBI), and a curated subset of the RefSeq experimental mRNA database (NCBI). Initial searches were performed on repeat-masked sequence with BLAST 2.0 (*54*) optimized for the Compaq Alpha compute-server and an effective database size of 3×10^9 for BLASTN searches and 1×10^9 for BLASTX searches. Additional processing of each query-subject pair was performed to improve the alignments. All protein BLAST results having an expectation score of $<1 \times 10^{-4}$, human nucleotide BLAST results having an expectation score of $<1 \times 10^{-8}$ with >94% identity, and rodent nucleotide BLAST results having an expectation score of $<1 \times 10^{8}$ with >80% identity were then examined on the basis of their high-scoring pair (HSP) coordinates on the scaffold to remove redundant hits, retaining hits that supported possible alternative splicing. For BLASTX searches, analysis was performed separately for selected model organisms (yeast, mouse, human, *C. elegans*, and *D. melanogaster*) so as not to exclude HSPs from these organisms that support the same gene structure. Sequences producing BLAST hits judged to be informative, nonredundant, and sufficiently similar to the scaffold sequence were then realigned to the genomic sequence with Sim4 for ESTs, and with Lap for proteins. Because both of these algorithms take splicing into account, the resulting alignments usually give a better representation of intron-exon boundaries than standard BLAST analyses and thus facilitate further annotation (both machine and human). In addition to the homology-based analysis described above, three ab initio gene prediction programs were used (*63*).
63. E. C. Uberbacher, Y. Xu, R. J. Mural, *Methods Enzymol.* **266**, 259 (1996); C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); R. J. Mural, *Methods Enzymol.* **303**, 77 (1999); A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); Floreal *et al.*, *Genome Res.* **8**, 967 (1998).
64. G. L. Miklos, B. John, *Am. J. Hum. Genet.* **31**, 264 (1979); U. Francke, *Cytogenet. Cell Genet.* **65**, 206 (1994).
65. P. E. Warburton, H. F. Willard, in *Human Genome Evolution*, M. S. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.
66. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000).
67. W. A. Bickmore, A. T. Sumner, *Trends Genet.* **5**, 144 (1989).
68. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).
69. G. Bernardi, *Gene* **241**, 3 (2000).
70. S. Zoubak, O. Clay, G. Bernardi, *Gene* **174**, 95 (1996).
71. S. Ohno, *Trends Genet.* **1**, 160 (1985).
72. K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, *Am. J. Hum. Genet.* **63**, 861 (1998).
73. M. J. McEachern, A. Krauskopf, E. H. Blackburn, *Annu. Rev. Genet.* **34**, 331 (2000).
74. A. Bird, *Trends Genet.* **3**, 342 (1987).
75. M. Gardiner-Garden, M. Frommer, *J. Mol. Biol.* **196**, 261 (1987).
76. F. Larsen, G. Gundersen, R. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).
77. S. H. Cross, A. Bird, *Curr. Opin. Genet. Dev.* **5**, 309 (1995).
78. J. Peters, *Genome Biol.* **1**, reviews1028.1 (2000) (http://genomebiology.com/2000/1/5/reviews/1028).
79. C. Grunau, W. Hindermann, A. Rosenthal, *Hum. Mol. Genet.* **9**, 2651 (2000).
80. F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11995 (1993).
81. S. H. Cross *et al.*, *Mamm. Genome* **11**, 373 (2000).
82. D. Slavov *et al.*, *Gene* **247**, 215 (2000).
83. A. F. Smit, A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).
84. D. J. Elliott *et al.*, *Hum. Mol. Genet.* **9**, 2117 (2000).
85. A. V. Makeyev, A. N. Chkheidze, S. A. Lievhaber, *J. Biol. Chem.* **274**, 24849 (1999).
86. Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craigen, *Genomics* **59**, 282 (1999).
87. P. Nouvel, *Genetica* **93**, 191 (1994).
88. I. Goncalves, L. Duret, D. Mouchiroud, *Genome Res.* **10**, 672 (2000).
89. Lek first compares all proteins in the proteome to one another. Next, the resulting BLAST reports are parsed, and a graph is created wherein each protein constitutes a node; any hit between two proteins with an expectation beneath a user-specified threshold constitutes an edge. Lek then uses this graph to compute a similarity between each protein pair *ij* in the context of the graph as a whole by simply dividing the number of BLAST hits shared in common between the two proteins by the total number of proteins hit by *i* and *j*. This simple metric has several interesting properties. First, because the similarity metric takes into account both the similarity and the differences between the two sequences at the level of BLAST hits, the metric respects the multidomain nature of protein space. Two multidomain proteins, for instance, each containing domains A and B, will have a greater pairwise similarity to each other than either one will have to a protein containing only A or B domains, so long as A-B–containing multidomain proteins are less frequent in the proteome than are single-domain proteins containing A or B domains. A second interesting property of this similarity metric is that it can be used to produce a similarity matrix for the proteome as a whole without having to first produce a multiple alignment for each protein family, an error-prone and very time-consuming process. Finally, the metric does not require that either sequence have significant homology to the other in order to have a defined similarity to each other, only that they

share at least one significant BLAST hit in common. This is an especially interesting property of the metric, because it allows the rapid recovery of protein families from the proteome for which no multiple alignment is possible, thus providing a computational basis for the extension of protein homology searches beyond those of current HMM- and profile-based search methods. Once the whole-proteome similarity matrix has been calculated, Lek first partitions the proteome into single-linkage clusters (*27*) on the basis of one or more shared BLAST hits between two sequences. Next, these single-linkage clusters are further partitioned into subclusters, each member of which shares a user-specified pairwise similarity with the other members of the cluster, as described above. For the purposes of this publication, we have focused on the analysis of single-linkage clusters and what we have termed "complete clusters," e.g., those subclusters for which every member has a similarity metric of 1 to every other member of the subcluster. We believe that the single-linkage and complete clusters are of special interest, in part, because they allow us to estimate and to compare sizes of core protein sets in a rigorous manner. The rationale for this is as follows: if one imagines for a moment a perfect clustering algorithm capable of perfectly partitioning one or more perfectly annotated protein sets into protein families, it is reasonable to assume that the number of clusters will always be greater than, or equal to, the number of single-linkage clusters, because single-linkage clustering is a maximally agglomerative clustering method. Thus, if there exists a single protein in the predicted protein set containing domains A and B, then it will be clustered by single linkage together with all single-domain proteins containing domains A or B. Likewise, for a predicted protein set containing a single multidomain protein, the number of real clusters must always be less than or equal to the number of complete clusters, because it is impossible to place a unique multidomain protein into a complete cluster. Thus, the single-linkage and complete clusters plus singletons should comprise a lower and upper bound of sizes of core protein sets, respectively, allowing us to compare the relative size and complexity of different organisms' predicted protein set.

90. T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
91. A. L. Delcher *et al.*, *Nucleic Acids Res.* **27**, 2369 (1999).
92. *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
93. The probability that a contiguous set of proteins is the result of a segmental duplication can be estimated approximately as follows. Given that protein A and B occur on one chromosome, and that A′ and B′ (paralogs of A and B) also exist in the genome, the probability that B′ occurs immediately after A′ is $1/N$, where N is the number of proteins in the set (for this analysis, $N = 26{,}588$). Allowing for B′ to occur as any of the next J-1 proteins [leaving a gap between A′ and B′ increases the probability to $(J - 1)/N$; allowing B′A′ or A′B′ gives a probability of $2(J - 1)/N$]. Considering three genes ABC, the probability of observing A′B′C′ elsewhere in the genome, given that the paralogs exist, is $1/N^2$. Three proteins can occur across a spread of five positions in six ways; more generally, we compute the number of ways that K proteins can be spread across J positions by counting all possible arrangements of $K - 2$ proteins in the $J - 2$ positions between the first and last protein. Allowing for a spread to vary from K positions (no gaps) to J gives

$$L = \sum_{X=K-2}^{J-2} \binom{X}{K-2}$$

arrangements. Thus, the probability of chance occurrence is L/N^{K-1}. Allowing for both sets of genes (e.g., ABC and A′B′C′) to be spread across J positions increases this to L^2/N^{K-1}. The duplicated segment might be rearranged by the operations of reversal or translocation; allowing for M such rearrangements gives us a probability $P = L^2M/N^{K-1}$. For example, the probability of observing a duplicated set of three genes in two different locations, where the three genes occur across a spread of five positions in both locations, is $36/N^2$; the expected number of such matched sets in the predicted protein set is approximately $(N)36/N^2 = 36/N$, a value $\ll 1$. Therefore, any such duplications of three genes are unlikely to result from random rearrangements of the genome. If any of the genes occur in more than two copies, the probability that the apparent duplication has occurred by chance increases. The algorithm for selecting candidate duplications only generates matched protein sets with $P \ll 1$..
94. B. J. Trask *et al.*, *Hum. Mol. Genet.* **7**, 13 (1998); D. Sharon *et al.*, *Genomics* **61**, 24 (1999).
95. W. B. Barbazuk *et al.*, *Genome Res.* **10**, 1351 (2000); A. McLysaght, A. J. Enright, L. Skrabanek, K. H. Wolfe, *Yeast* **17**, 22 (2000); D. W. Burt *et al.*, *Nature* **402**, 411 (1999).
96. Reviewed in L. Skrabanek, K. H. Wolfe, *Curr. Opin. Genet. Dev.* **8**, 694 (1998).
97. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, *Genome Res.* **8**, 748 (1998); P. Taillon-Miller, E. E. Piemot, P. Y. Kwok, *Genome Res.* **9**, 499 (1999).
98. D. Altshuler *et al.*, *Nature* **407**, 513 (2000).
99. G. T. Marth *et al.*, *Nature Genet.* **23**, 452 (1999).
100. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
101. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
102. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
103. J. Zhang, T. L. Madden, *Genome Res.* **7**, 649 (1997).
104. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
105. From the observed coverage of the sequences at each site for each individual, we calculated the probability that a SNP would be detected at the site if it were present. For each level of coverage, there is a binomial sampling of the two homologs for each individual, and a heterozygous site could only be ascertained if both homologs are present, or if two alleles from different individuals are present. With coverage x from a given individual, both homologs are present in the assembly with probability $1 - (1/2)x^{-1}$. Even if both homologs are present, the probability that a SNP is detected is <1 because a fraction of sites failed the quality criteria. Integrating over coverage levels, the binomial sampling, and the quality distribution, we derived an expected number of sites in the genome that were ascertained for polymorphism for each individual. The nucleotide diversity was then the observed number of variable sites divided by the expected number of sites ascertained.
106. M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, *Genetics* **150**, 1133 (1998).
107. D. A. Nickerson *et al.*, *Nature Genet.* **19**, 233 (1998); D. A. Nickerson *et al.*, *Genomic Res.* **10**, 1532 (2000); L. Jorde *et al.*, *Am. J. Hum. Genet.* **66**, 979 (2000); D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
108. M. Przeworski, R. R. Hudson, A. Di Rienzo, *Trends Genet.* **16**, 296 (2000).
109. S. Tavare, *Theor. Popul. Biol.* **26**, 119 (1984).
110. R. R. Hudson, in *Oxford Surveys in Evolutionary Biology*, D. J. Futuyma, J. D. Antonovics, Eds. (Oxford Univ. Press, Oxford, 1990), vol. 7, pp. 1–44.
111. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
112. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
113. H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, *Nature Genet.* **22**, 78 (1999).
114. E. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
115. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
116. Brief description of the methods used to build the Panther classification. First, the June 2000 release of the GenBank NR protein database (excluding sequences annotated as fragments or mutants) was partitioned into clusters using BLASTP. For the clustering, a seed sequence was randomly chosen, and the cluster was defined as all sequences matching the seed to statistical significance (E-value $< 10^{-5}$) and "globally" alignable (the length of the match region must be >70% and <130% of the length of the seed). If the cluster had more than five members, and at least one from a multicellular eukaryote, the cluster was extended. For the extension step, a hidden Markov Model (HMM) was trained for the cluster, using the SAM software package, version 2..The HMM was then scored against GenBank NR (excluding mutants but including fragments for this step), and all sequences scoring better than a specific (NLL-NULL) score were added to the cluster. The HMM was then retrained (with fixed model length) and all sequences in the cluster were aligned to the HMM to produce a multiple sequence alignment. This alignment was assessed by a number of quality measures. If the alignment failed the quality check, the initial cluster was rebuilt around the seed using a more restrictive E-value, followed by extension, alignment, and reassessment. This process was repeated until the alignment quality was good. The multiple alignment and "general" (i.e., describing the entire cluster, or "family") HMM (*176*) were then used as input into the BETE program (*177*). BETE calculates a phylogenetic tree for the sequences in the alignment. Functional information about the sequences in each cluster were parsed from SwissProt (*178*) and GenBank records. "Tree-attribute viewer" software was used by biologist curators to correlate the phylogenetic tree with protein function. Subfamilies were manually defined on the basis of shared function across subtrees, and were named accordingly. HMMs were then built for each subfamily, using information from both the subfamily and family (K. Sjölander, in preparation). Families were also manually named according to the functions contained within them. Finally, all of the families and subfamilies were classified into categories and subcategories based on their molecular functions. The categorization was done by manual review of the family and subfamily names, by examining SwissProt and GenBank records, and by review of the literature as well as resources on the World Wide Web. The current version (2.0) of the Panther molecular function schema has four levels: category, subcategory, family, and subfamily. Protein sequences for whole eukaryotic genomes (for the predicted human proteins and annotated proteins for fly, worm, yeast, and *Arabidopsis*) were scored against the Panther library of family and subfamily HMMs. If the score was significant (the NLL-NULL score cutoff depends on the protein family), the protein was assigned to the family or subfamily function with the most significant score.
117. C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, *Nucleic Acids Res.* **27**, 229 (1999).
118. A. Goffeau *et al.*, *Science* **274**, 546, 563 (1996).
119. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
120. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
121. E. R. Kandel, J. H. Schwartz, T. Jessell, *Principles of Neural Science* (McGraw-Hill, New York, ed. 4, 2000).
122. D. A. Goodenough, J. A. Goliger, D. L. Paul, *Annu. Rev. Biochem.* **65**, 475 (1996).
123. D. G. Wilkinson, *Int. Rev. Cytol.* **196**, 177 (2000).
124. F. Nakamura, R. G. Kalb, S. M. Strittmatter, *J. Neurobiol.* **44**, 219 (2000).
125. P. J. Horner, F. H. Gage, *Nature* **407**, 963 (2000); P. Casaccia-Bonnefil, C. Gu, M. V. Chao, *Adv. Exp. Med. Biol.* **468**, 275 (1999).
126. S. Wang, B. A. Barres, *Neuron* **27**, 197 (2000).
127. M. Geppert, T. C. Sudhof, *Annu. Rev. Neurosci.* **21**, 75 (1998); J. T. Littleton, H. J. Bellen, *Trends Neurosci.* **18**, 177 (1995).
128. A. Maximov, T. C. Sudhof, I. Bezprozvanny, *J. Biol. Chem.* **274**, 24453 (1999).
129. B. Sampo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3666 (2000).
130. G. Lemke, *Glia* **7**, 263 (1993).
131. M. Bernfield *et al.*, *Annu. Rev. Biochem.* **68**, 729 (1999).
132. N. Perrimon, M. Bernfield, *Nature* **404**, 725 (2000).
133. U. Lindahl, M. Kusche-Gullberg, L. Kjellen, *J. Biol. Chem.* **273**, 24979 (1998).
134. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
135. T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, *J. Biol. Chem.* **274**, 25555 (1999).

136. R. A. Black, J. M. White, *Curr. Opin. Cell Biol.* **10**, 654 (1998).

137. L. Aravind, V. M. Dixit, E. V. Koonin, *Trends Biochem. Sci.* **24**, 47 (1999).

138. A. G. Uren *et al.*, *Mol. Cell* **6**, 961 (2000).

139. P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, *Mamm. Genome* **4**, 695 (1993).

140. K. Meyer-Siegler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8460 (1991).

141. N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, *Nucleic Acids Res.* **21**, 993 (1993).

142. N. A. Tatton, *Exp. Neurol.* **166**, 29 (2000).

143. N. Kenmochi *et al.*, *Genome Res.* **8**, 509 (1998).

144. F. W. Chen, Y. A. Ioannou, *Int. Rev. Immunol.* **18**, 429 (1999).

145. H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, *Nucleic Acids Res.* **18**, 1513 (1990).

146. D. M. Chambers, J. Peters, C. M. Abbott, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4463 (1998); A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, *Dev. Dyn.* **216**, 267 (1999).

147. D. Aeschlimann, V. Thomazy, *Connect. Tissue Res.* **41**, 1 (2000).

148. P. Munroe *et al.*, *Nature Genet.* **21**, 142 (1999); S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, *Science* **254**, 1634 (1991); B. Furie *et al.*, *Blood* **93**, 1798 (1999).

149. J. W. Kehoe, C. R. Bertozzi, *Chem. Biol.* **7**, R57 (2000).

150. T. Pawson, P. Nash, *Genes Dev.* **14**, 1027 (2000).

151. A. W. van der Velden, A. A. Thomas, *Int. J. Biochem. Cell Biol.* **31**, 87 (1999).

152. C. M. Fraser *et al.*, *Science* **281**, 375 (1998); H. Tettelin *et al.*, *Science* **287**, 1809 (2000).

153. D. Brett *et al.*, *FEBS Lett.* **474**, 83 (2000).

154. H. J. Muller, H. Kern, *Z. Naturforsch. B* **22**, 1330 (1967).

155. H. J. Muller, in *Heritage from Mendel*, R. A. Brink, Ed. (Univ. of Wisconsin Press, Madison, WI, 1967), p. 419.

156. J. F. Crow, M. Kimura, *Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).

157. K. Kobayashi *et al.*, *Nature* **394**, 388 (1998).

158. A. P. Feinberg, *Curr. Top. Microbiol. Immunol.* **249**, 87 (2000).

159. C. A. Collins, C. Guthrie, *Nature Struct. Biol.* **7**, 850 (2000).

160. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999).

161. Q. Wang, J. Khillan, P. Gadue, K. Nishikura, *Science* **290**, 1765 (2000).

162. M. Holcik, N. Sonenberg, R. G. Korneluk, *Trends Genet.* **16**, 469 (2000).

163. T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, *Nature* **408**, 106 (2000).

164. E. Capanna, M. G. M. Romanini, *Caryologia* **24**, 471 (1971).

165. J. Maynard Smith, *J. Theor. Biol.* **128**, 247 (1987).

166. D. Charlesworth, B. Charlesworth, M. T. Morgan, *Genetics* **141**, 1619 (1995).

167. J. E. Bailey, *Nature Biotechnol.* **17**, 616 (1999).

168. R. Maleszka, H. G. de Couet, G. L. Miklos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3731 (1998).

169. G. L. Miklos, *J. Neurobiol.* **24**, 842 (1993).

170. J. P. Crutchfield, K. Young, *Phys. Rev. Lett.* **63**, 105 (1989); M. Gell-Mann, S. Lloyd, *Complexity* **2**, 44 (1996).

171. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).

172. E. Colucci-Guyon *et al.*, *Cell* **79**, 679 (1994).

173. J. Sambrook, E. F. Fritch, T. Maniatis, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 1989).

174. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998); B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).

175. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).

176. A. Krogh, K. Sjölander, *J. Mol. Biol.* **235**, 1501 (1994).

177. K. Sjölander, *Proc. Int. Soc. Mol. Biol.* **6**, 165 (1998).

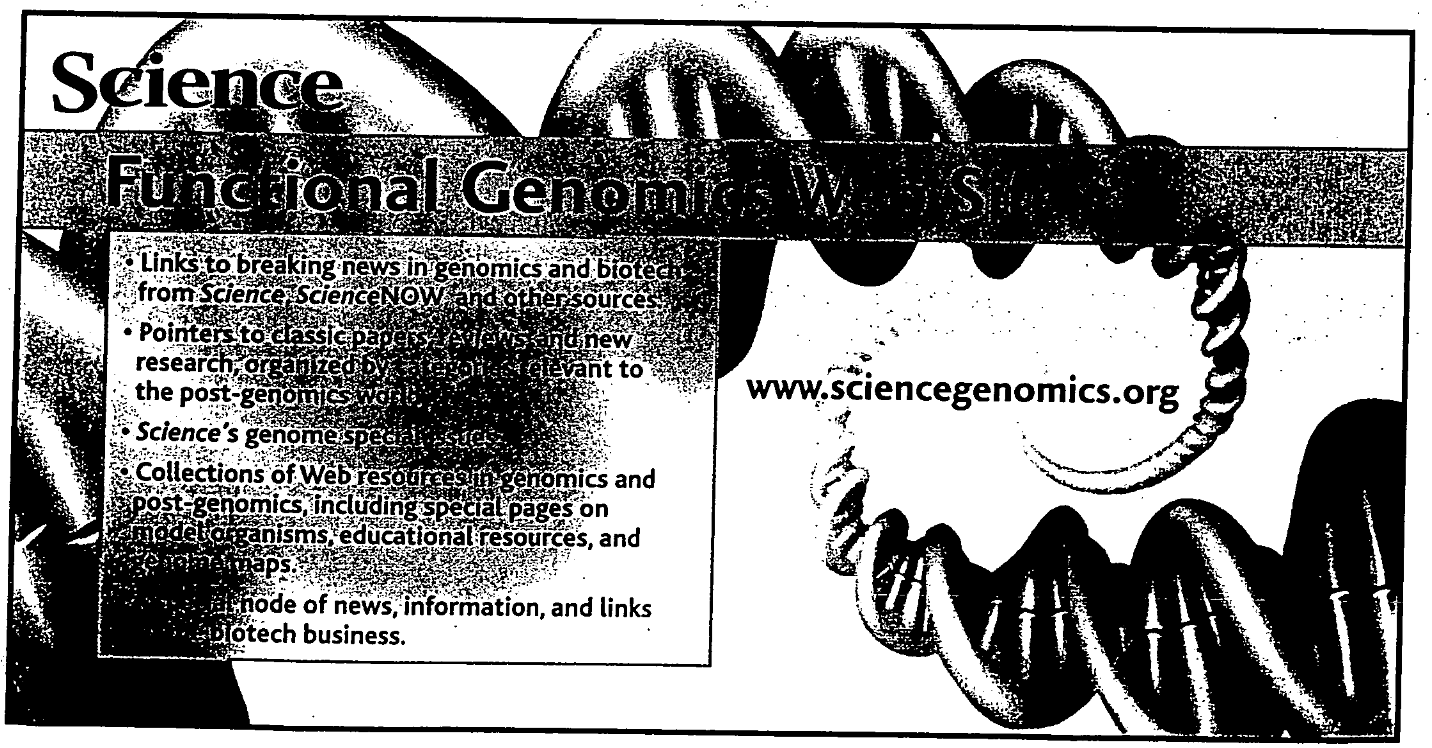
178. A. Bairoch, R. Apweiler, *Nucleic Acids Res.* **28**, 45 (2000).

179. GO, available at www.geneontology.org/.

180. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, *Nucleic Acids Res.* **28**, 33 (2000).

181. We thank E. Eichler and J. L. Goldstein for many helpful discussions and critical reading of the manuscript, and A. Caplan for advice and encouragement. We also thank T. Hein, D. Lucas, G. Edwards, and the Celera IT staff for outstanding computational support. The cost of this project was underwritten by the Celera Genomics Group of the Applera Corporation. We thank the Board of Directors of Applera Corporation: J. F. Abely Jr. (retired), R. H. Ayers, J.-L. Bélingard, R. H. Hayes, A. J. Levine, T. E. Martin, C. W. Slayman, O. R. Smith, G. C. St. Laurent Jr., and J. R. Tobin for their vision, enthusiasm, and unwavering support and T. L. White for leadership and advice. Data availability: The genome sequence and additional supporting information are available to academic scientists at the Web site (www.celera.com). Instructions for obtaining a DVD of the genome sequence can be obtained through the Web site. For commercial scientists wishing to verify the results presented here, the genome data are available upon signing a Material Transfer Agreement, which can also be found on the Web site.

5 December 2000; accepted 19 January 2001



DD

THE HUMAN GENOME

Humanity has been given a great gift. With the completion of the human genome sequence, we have received a powerful tool for unlocking the secrets of our genetic heritage and for finding our place among the other participants in the adventure of life.

This week's issue of *Science* contains the report of the sequencing of the human genome from a group of authors led by Craig Venter of Celera Genomics. The report of the sequencing of the human genome from the publicly funded consortium of laboratories led by Francis Collins appears in this week's *Nature*. This stunning achievement has been portrayed—often unfairly—as a competition between two ventures, one public and one private. That characterization detracts from the awesome accomplishment jointly unveiled this week. In truth, each project contributed to the other. The inspired vision that launched the publicly funded project roughly 10 years ago reflected, and now rewards, the confidence of those who believe that the pursuit of large-scale fundamental problems in the life sciences is in the national interest. The technical innovation and drive of Craig Venter and his colleagues made it possible to celebrate this accomplishment far sooner than was believed possible. Thus, we can salute what has become, in the end, not a contest but a marriage (perhaps encouraged by shotgun) between public funding and private entrepreneurship.

A historic moment for the scientific endeavor.

There are excellent scientific reasons for applauding an outcome that has given us two winners. Two sequences are better than one; the opportunity for comparison and convergence is invaluable. Indeed, a real-world proof of the importance of access to both sets of data can be found in the pages of this issue of *Science*, in the comparative analysis by Olivier *et al.* (p. 1298).

Although we have made the point before, it is worth repeating that the sequencing of the human genome represents, not an ending, but the beginning of a new approach to biology. As Galas says in his Viewpoint (p. 1257), the knowledge that all of the genetic components of any process can be identified will give extraordinary new power to scientists. Because of this breakthrough, research can evolve from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. Several articles in this issue highlight how this approach is already beginning to revolutionize the way we look at human disease.

This has been a massive project, on a scale unparalleled in the history of biology, but of course it has built on the scientific insights of centuries of investigators. By coincidence, this landmark announcement falls during the week of the anniversary of the birth of Charles Darwin. Darwin's message that the survival of a species can depend on its ability to evolve in the face of change is peculiarly pertinent to discussions that have gone on in the past year over access to the Celera data. (Full information regarding the agreements that were reached to make the data available can be found at www.sciencemag.org/feature/data/announcement/gsp.shl.) We are willing to be flexible in allowing data repositories other than the traditional GenBank, while insisting on access to all the data needed to verify conclusions. In this domain, change is everywhere. Commercial researchers are producing more and more potentially valuable sequences, yet (at least in the United States) laws governing databases provide scant protection against piracy. Had the Celera data been kept secret, it would have been a serious loss to the scientific community. We hope that our adaptability in the face of change will enable other proprietary data to be published after peer review, in a way that satisfies our continuing commitment to full access.

It should be no surprise that an achievement so stunning, and so carefully watched, has created new challenges for the scientific venture. *Science* is proud to have played a role in bringing this discovery onto the public stage. It is literally true that this is a historic moment for the scientific endeavor. The human genome has been called the Book of Life. Rather, it is a library, in which, with rules that encourage exploration and reward creativity, we can find many of the books that will help define us and our place in the great tapestry of life.

Barbara R. Jasny and Donald Kennedy

EE

```
Query= SEQ ID NO:1
         (3924 letters)

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

AC097715.3.1.143642                                               526   e-146
AC019105.7.1.170491                                               482   e-133
AC019159.8.1.163085                                               436   e-119
AC104648.2.1.112084                                               401   e-108
AC074362.5.1.149705                                               387   e-104
AC079154.5.1.175387                                               163   1e-36

>AC097715.3.1.143642
          Length = 143642

 Score =  526 bits (265), Expect = e-146
 Identities = 266/267 (99%)
 Strand = Plus / Plus


Query: 1062   gggcaatgtcactttttcctgctccgaaccacagattgtgcccatcacatttgtyaactc 1121
              ||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct: 21277 gggcaatgtcactttttcctgctccgaaccacagattgtgcccatcacatttgtcaactc 21336


Query: 1122   cagcggcagctatttgctgctgcccggcacccccaaattgatgggctctcagtgagttt 1181
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 21337 cagcggcagctatttgctgctgcccggcacccccaaattgatgggctctcagtgagttt 21396


Query: 1182   ccagtttcgaacatggaacaaggatggtctgcttctgtccacagagctgtctgagggctc 1241
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 21397 ccagtttcgaacatggaacaaggatggtctgcttctgtccacagagctgtctgagggctc 21456


Query: 1242   gggaaccctgctgctgagcctggagggtggaatcctgagactcgtgattcagaaaatgac 1301
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 21457 gggaaccctgctgctgagcctggagggtggaatcctgagactcgtgattcagaaaatgac 21516


Query: 1302   agaacgcgtagctgaaatcctcacagg 1328
              |||||||||||||||||||||||||||
Sbjct: 21517 agaacgcgtagctgaaatcctcacagg 21543



 Score =  349 bits (176), Expect = 7e-93
 Identities = 176/176 (100%)
 Strand = Plus / Plus


Query: 1476   agggtgccccgacaatctcaccgattcccaatgtttaaatcccattaaggctttccaagg 1535
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44269 agggtgccccgacaatctcaccgattcccaatgtttaaatcccattaaggctttccaagg 44328
```

```
Query: 1536  ctgcatgaggctcatctttattgataaccagcccaaggacctcatttcagttcagcaagg 1595
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44329 ctgcatgaggctcatctttattgataaccagcccaaggacctcatttcagttcagcaagg 44388


Query: 1596  ttccctggggaattttagtgatttacacattgatctgtgtagcatcaaagacaggt 1651
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44389 ttccctggggaattttagtgatttacacattgatctgtgtagcatcaaagacaggt 44444


 Score =  305 bits (154), Expect = 1e-79
 Identities = 154/154 (100%)
 Strand = Plus / Plus


Query: 1325  caggcagcaacttgaatgatggcctgtggcactcggttagcatcaacgccaggaggaacc 1384
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 41286 caggcagcaacttgaatgatggcctgtggcactcggttagcatcaacgccaggaggaacc 41345


Query: 1385  gcatcacgctcactctggatgatgaagcagcacccccggctccagacagcacttgggtgc 1444
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 41346 gcatcacgctcactctggatgatgaagcagcacccccggctccagacagcacttgggtgc 41405


Query: 1445  agatttattctggaaatagctactattttggagg 1478
             ||||||||||||||||||||||||||||||||||
Sbjct: 41406 agatttattctggaaatagctactattttggagg 41439


 Score =  238 bits (120), Expect = 2e-59
 Identities = 120/120 (100%)
 Strand = Plus / Plus


Query: 1757   ccatctacgagcaatcctgcgaggtgtacaggcaccaggggaatacagccggcttcttct 1816
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 126787 ccatctacgagcaatcctgcgaggtgtacaggcaccaggggaatacagccggcttcttct 126846


Query: 1817   acatcgactcagatggcagcggcccactgggacctctccaggtgtactgcaatatcactg 1876
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 126847 acatcgactcagatggcagcggcccactgggacctctccaggtgtactgcaatatcactg 126906


 Score =  216 bits (109), Expect = 7e-53
 Identities = 109/109 (100%)
 Strand = Plus / Plus


Query: 1648  aggtgtttgccaaactactgtgaacatggaggaagctgctcccagtcctggactaccttc 1707
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 80201 aggtgtttgccaaactactgtgaacatggaggaagctgctcccagtcctggactaccttc 80260
```

```
Query: 1708  tattgtaactgcagtgacacaagttacactggtgccacctgccacaact 1756
             |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 80261 tattgtaactgcagtgacacaagttacactggtgccacctgccacaact 80309


>AC019105.7.1.170491
          Length = 170491

 Score =  482 bits (243), Expect = e-133
 Identities = 243/243 (100%)
 Strand = Plus / Plus


Query: 2750 taggggaacgtcatccagacagaaaggcttcctaggatgcattcgctccttacacttga 2809
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 152  taggggaacgtcatccagacagaaaggcttcctaggatgcattcgctccttacacttga 211


Query: 2810 atggacagaaaatggacctggaagagagggcaaaggtcacatctggagtcaggccaggct 2869
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 212  atggacagaaaatggacctggaagagagggcaaaggtcacatctggagtcaggccaggct 271


Query: 2870 gccccggccactgcagcagctacggcagcatctgccacaacgggggcaagtgtgtggaga 2929
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 272  gccccggccactgcagcagctacggcagcatctgccacaacgggggcaagtgtgtggaga 331


Query: 2930 agcacaatggctacctgtgtgattgcaccaattcaccttatgaagggccctttgcaaaa 2989
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 332  agcacaatggctacctgtgtgattgcaccaattcaccttatgaagggccctttgcaaaa 391


Query: 2990 aag 2992
            |||
Sbjct: 392  aag 394


 Score =  452 bits (228), Expect = e-124
 Identities = 228/228 (100%)
 Strand = Plus / Plus


Query: 2991 agaggtttctgctgtttttgaggctggcacgtcggttacttacatgtttcaagaacccta 3050
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8347 agaggtttctgctgtttttgaggctggcacgtcggttacttacatgtttcaagaacccta 8406


Query: 3051 tcctgtgaccaagaatataagcctctcatcctcagctatttacacagattcagctccatc 3110
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8407 tcctgtgaccaagaatataagcctctcatcctcagctatttacacagattcagctccatc 8466
```

```
Query: 3111 caaggaaaacattgcacttagctttgtgacaacccaggcacccagtcttttgctctttat 3170
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8467 caaggaaaacattgcacttagctttgtgacaacccaggcacccagtcttttgctctttat 8526


Query: 3171 caattcttcttctcaggacttcgtggttgttctgctctgcaagaatgg 3218
            ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8527 caattcttcttctcaggacttcgtggttgttctgctctgcaagaatgg 8574


 Score =  440 bits (222), Expect = e-120
 Identities = 222/222 (100%)
 Strand = Plus / Plus


Query: 3434   cagagaatcttggtttggattctgaagttgctaaagcaaatgccatgggttttgctggat 3493
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113132 cagagaatcttggtttggattctgaagttgctaaagcaaatgccatgggttttgctggat 113191


Query: 3494   gcatgtcttccgtccagtacaaccacatagcaccactgaaggctgccctgcgccatgcca 3553
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113192 gcatgtcttccgtccagtacaaccacatagcaccactgaaggctgccctgcgccatgcca 113251


Query: 3554   ctgtcgcgcctgtgactgtccatgggaccttgacggaatccagctgtggcttcatggtgg 3613
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113252 ctgtcgcgcctgtgactgtccatgggaccttgacggaatccagctgtggcttcatggtgg 113311


Query: 3614   actcagatgtgaatgcagtgaccacggtgcattcttcatcag 3655
              ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113312 actcagatgtgaatgcagtgaccacggtgcattcttcatcag 113353


 Score =  395 bits (199), Expect = e-106
 Identities = 199/199 (100%)
 Strand = Plus / Plus


Query: 3726   aggggtgatagcagtggtgatattcatcatcttctgtatcatcggcatcatgacccggtt 3785
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 124343 aggggtgatagcagtggtgatattcatcatcttctgtatcatcggcatcatgacccggtt 124402


Query: 3786   cctctaccagcacaagcagtcacatcgtacgagccagatgaaggagaaggaatatccaga 3845
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 124403 cctctaccagcacaagcagtcacatcgtacgagccagatgaaggagaaggaatatccaga 124462


Query: 3846   aaatttggacagttccttcagaaatgaaattgacttgcaaaacacagtgagcgagtgtaa 3905
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 124463 aaatttggacagttccttcagaaatgaaattgacttgcaaaacacagtgagcgagtgtaa 124522
```

```
Query: 3906   acgggaatatttcatctga 3924
              |||||||||||||||||||
Sbjct: 124523 acgggaatatttcatctga 124541


 Score =  262 bits (132), Expect = 1e-66
 Identities = 132/132 (100%)
 Strand = Plus / Plus


Query: 3217  ggaagcttacaggttcgctatcacctaaacaaggaagaaacccatgtattcaccattgat 3276
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 75558 ggaagcttacaggttcgctatcacctaaacaaggaagaaacccatgtattcaccattgat 75617


Query: 3277  gcagataactttgctaacagaaggatgcaccacttgaagattaaccgagagggaagagag 3336
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 75618 gcagataactttgctaacagaaggatgcaccacttgaagattaaccgagagggaagagag 75677


Query: 3337  cttaccattcag 3348
             ||||||||||||
Sbjct: 75678 cttaccattcag 75689


 Score =  180 bits (91), Expect = 4e-42
 Identities = 91/91 (100%)
 Strand = Plus / Plus


Query: 3346  cagatggaccagcaacttcgactcagttataacttctctccggaagtagagttcagggtt 3405
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 79925 cagatggaccagcaacttcgactcagttataacttctctccggaagtagagttcagggtt 79984


Query: 3406  ataaggtcactcaccttgggcaaagtcacag 3436
             |||||||||||||||||||||||||||||||
Sbjct: 79985 ataaggtcactcaccttgggcaaagtcacag 80015


 Score =  153 bits (77), Expect = 1e-33
 Identities = 77/77 (100%)
 Strand = Plus / Plus


Query: 3652   tcagatccttttgggaagacagatgagcgggaaccactcacaaatgctgttcgaagtgat 3711
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121716 tcagatccttttgggaagacagatgagcgggaaccactcacaaatgctgttcgaagtgat 121775


Query: 3712   tcggcagtcatcggagg 3728
              |||||||||||||||||
Sbjct: 121776 tcggcagtcatcggagg 121792
```

```
>AC019159.8.1.163085
          Length = 163085

 Score =  436 bits (220), Expect = e-119
 Identities = 220/220 (100%)
 Strand = Plus / Plus


Query: 2534   ctccttcagagatcacctttgccatcgatgttgggaatggtcctgtggagcttgtagtcc 2593
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 148143 ctccttcagagatcacctttgccatcgatgttgggaatggtcctgtggagcttgtagtcc 148202


Query: 2594   agtctccttctcttctgaatgacaaccaatggcactatgtccgggctgagaggaacctca 2653
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 148203 agtctccttctcttctgaatgacaaccaatggcactatgtccgggctgagaggaacctca 148262


Query: 2654   aggagacctccctgcaggtggacaaccttccaaggagcaccagggagacgtcggaggagg 2713
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 148263 aggagacctccctgcaggtggacaaccttccaaggagcaccagggagacgtcggaggagg 148322


Query: 2714   gccattttcgactgcagctgaacagccagttgtttgtagg 2753
               ||||||||||||||||||||||||||||||||||||||||
Sbjct: 148323 gccattttcgactgcagctgaacagccagttgtttgtagg 148362



 Score =  401 bits (202), Expect = e-108
 Identities = 202/202 (100%)
 Strand = Plus / Plus


Query: 1876  gaggacaagatctggacatcagtgcagcacaacaatacagagctgacccgagtgcggggc 1935
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 23101 gaggacaagatctggacatcagtgcagcacaacaatacagagctgacccgagtgcggggc 23160


Query: 1936  gctaaccctgagaagccctatgccatggccttggactacgggggcagcatggaacagctg 1995
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 23161 gctaaccctgagaagccctatgccatggccttggactacgggggcagcatggaacagctg 23220


Query: 1996  gaggccgtgatcgacggctctgagcactgtgagcaggaggtggcctaccactgcaggagg 2055
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 23221 gaggccgtgatcgacggctctgagcactgtgagcaggaggtggcctaccactgcaggagg 23280


Query: 2056  tcccgcctgctcaacacgccgg 2077
              ||||||||||||||||||||||
Sbjct: 23281 tcccgcctgctcaacacgccgg 23302
```

```
 Score =  339 bits (171), Expect = 7e-90
 Identities = 171/171 (100%)
 Strand = Plus / Plus


Query: 2363   gacgcttctggaacgccgtctcattttatacagaagcctcttacctccactttcctacct 2422
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 139321 gacgcttctggaacgccgtctcattttatacagaagcctcttacctccactttcctacct 139380


Query: 2423   tccatgcggaattcagtgccgatatttccttctttttaaaaccacagcattatccggag 2482
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 139381 tccatgcggaattcagtgccgatatttccttctttttaaaaccacagcattatccggag 139440


Query: 2483   ttttcctagaaaatcttggcattaaagacttcattcgactcgaaataagct 2533
              |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 139441 ttttcctagaaaatcttggcattaaagacttcattcgactcgaaataagct 139491


 Score =  315 bits (159), Expect = 1e-82
 Identities = 159/159 (100%)
 Strand = Plus / Plus


Query: 2077   gatggaacaccatttacctggtggattgggcggtccaatgaaaggcacccttactgggga 2136
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 122572 gatggaacaccatttacctggtggattgggcggtccaatgaaaggcacccttactgggga 122631


Query: 2137   ggttcccctcctggggtccagcagtgtgagtgtggcctagacgagagctgcctggacatt 2196
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 122632 ggttcccctcctggggtccagcagtgtgagtgtggcctagacgagagctgcctggacatt 122691


Query: 2197   cagcacttttgcaattgcgacgctgacaaggatgaatgg 2235
              |||||||||||||||||||||||||||||||||||||||
Sbjct: 122692 cagcacttttgcaattgcgacgctgacaaggatgaatgg 122730


 Score =  258 bits (130), Expect = 2e-65
 Identities = 130/130 (100%)
 Strand = Plus / Plus


Query: 2234   ggacaaatgatactggctttctttccttcaaagaccacttgcctgtcactcagatagtta 2293
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 139015 ggacaaatgatactggctttctttccttcaaagaccacttgcctgtcactcagatagtta 139074


Query: 2294   tcactgataccgacagatcaaactcagaagccgcttggagaattggtcccttgcgttgct 2353
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 139075 tcactgataccgacagatcaaactcagaagccgcttggagaattggtcccttgcgttgct 139134
```

```
Query: 2354   atggtgaccg 2363
              ||||||||||
Sbjct: 139135 atggtgaccg 139144


>AC104648.2.1.112084
          Length = 112084

 Score =  401 bits (202), Expect = e-108
 Identities = 205/206 (99%)
 Strand = Plus / Plus


Query: 530   aatcagacgttgctgactttgatggccgaagctcacttctgtacaggttcaatcagaagt 589
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61554 aatcagatgttgctgactttgatggccgaagctcacttctgtacaggttcaatcagaagt 61613


Query: 590   tgatgagtactctcaaagatgtgatctccctgaagttcaagagcatgcaaggagatgggg 649
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61614 tgatgagtactctcaaagatgtgatctccctgaagttcaagagcatgcaaggagatgggg 61673


Query: 650   tcctgttccatggagaaggtcagcgtggagaccacatcaccttggaactccagaagggga 709
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61674 tcctgttccatggagaaggtcagcgtggagaccacatcaccttggaactccagaagggga 61733


Query: 710   ggctcgccctacacctcaatttgggt 735
             ||||||||||||||||||||||||||
Sbjct: 61734 ggctcgccctacacctcaatttgggt 61759



 Score =  365 bits (184), Expect = 1e-97
 Identities = 185/186 (99%)
 Strand = Plus / Plus


Query: 733   ggtgacagcaaagcgcggctcagcagcagcttgccctctgccaccctgggcagcctcctg 792
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 73822 ggtgacagcaaagcgcggctcagcagcagcttgccctctgccaccctgggcagcctcctg 73881


Query: 793   gatgaccagcactggcactyggtcctcattgagcgggtgggcaagcaggtgaacttcacg 852
             ||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct: 73882 gatgaccagcactggcactcggtcctcattgagcgggtgggcaagcaggtgaacttcacg 73941


Query: 853   gtggacaagcacacacagcacttccgcaccaagggcgagacggatgccttagacattgac 912
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 73942 gtggacaagcacacacagcacttccgcaccaagggcgagacggatgccttagacattgac 74001
```

```
Query: 913   tatgag 918
             ||||||
Sbjct: 74002 tatgag 74007


 Score =  296 bits (149), Expect = 1e-76
 Identities = 149/149 (100%)
 Strand = Plus / Plus


Query: 381   gacctttgcaggaaacatgaatgctgacagcgtggtgcaccacaagctattgcactcagt 440
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44512 gacctttgcaggaaacatgaatgctgacagcgtggtgcaccacaagctattgcactcagt 44571


Query: 441   gagagcccgatttgttcgctttgtgcccctggaatggaatcccagtgggaagattggcat 500
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 44572 gagagcccgatttgttcgctttgtgcccctggaatggaatcccagtgggaagattggcat 44631


Query: 501   gagagtcgaggtctacggatgttcctata 529
             |||||||||||||||||||||||||||||
Sbjct: 44632 gagagtcgaggtctacggatgttcctata 44660


 Score = 288 bits (145), Expect = 2e-74
 Identities = 146/147 (99%)
 Strand = Plus / Plus


Query: 917    agcttagttttggaggaattccagtaccaggaaaacctgggacctttttaaagaaaaact 976
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 101807 agcttagttttggaggaattccagtaccaggaaaacctgggacctttttaaagaaaaact 101866


Query: 977    tccatggatgcatcgaaaacctttactacaatggagtaaacataattracctggctaaga 1036
              ||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct: 101867 tccatggatgcatcgaaaacctttactacaatggagtaaacataattgacctggctaaga 101926


Query: 1037   gacgaaagcatcagatctatactgtgg 1063
              |||||||||||||||||||||||||||
Sbjct: 101927 gacgaaagcatcagatctatactgtgg 101953
```

```
>AC074362.5.1.149705
          Length = 149705

 Score =  387 bits (195), Expect = e-104
 Identities = 195/195 (100%)
 Strand = Plus / Plus


Query: 187   ggaactggcggttggtccccagcagattccaatgctcaacagtggctccagatggacctg 246
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 93165 ggaactggcggttggtccccagcagattccaatgctcaacagtggctccagatggacctg 93224


Query: 247   ggaaacagagtagagattacagcagtggccacgcagggaagatacggaagctctgactgg 306
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 93225 ggaaacagagtagagattacagcagtggccacgcagggaagatacggaagctctgactgg 93284


Query: 307   gtgacgagttacagcctgatgttcagtgacacaggacgcaactggaaacagtacaaacaa 366
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 93285 gtgacgagttacagcctgatgttcagtgacacaggacgcaactggaaacagtacaaacaa 93344


Query: 367   gaagacagcatctgg 381
              |||||||||||||||
Sbjct: 93345 gaagacagcatctgg 93359


 Score =  210 bits (106), Expect = 5e-51
 Identities = 106/106 (100%)
 Strand = Plus / Plus


Query: 83    acaactgtgatgatccactagcatccctgctctctccaatggctttttccagttcctcag 142
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 72671 acaactgtgatgatccactagcatccctgctctctccaatggctttttccagttcctcag 72730


Query: 143   acctcactggcactcacagcccagctcaactcaactggagagttgg 188
              ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 72731 acctcactggcactcacagcccagctcaactcaactggagagttgg 72776


>AC079154.5.1.175387
          Length = 175387

 Score =  163 bits (82), Expect = 1e-36
 Identities = 82/82 (100%)
 Strand = Plus / Plus


Query: 1      atggattctttaccacggctgaccagcgttttgactttgctgttctctggcttgtggcat 60
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 124330 atggattctttaccacggctgaccagcgttttgactttgctgttctctggcttgtggcat 124389
```

```
Query: 61     ttaggattaacagcgacaaact 82
              ||||||||||||||||||||||
Sbjct: 124390 ttaggattaacagcgacaaact 124411
```

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Boo

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

□ 1: AC097715. Homo sapiens BAC ...[gi:18056738]

Links

```
LOCUS       AC097715               143642 bp    DNA     linear   PRI 21-FEB-2002
DEFINITION  Homo sapiens BAC clone RP11-563A13 from 2, complete sequence.
ACCESSION   AC097715 AC027111
VERSION     AC097715.3  GI:18056738
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 143642)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
  MEDLINE   99063792
   PUBMED   9847074
REFERENCE   2  (bases 1 to 143642)
  AUTHORS   Tomlinson,C. and Kozlowicz,A.
  TITLE     The sequence of Homo sapiens BAC clone RP11-563A13
  JOURNAL   Unpublished (2001)
REFERENCE   3  (bases 1 to 143642)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-OCT-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 143642)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-JAN-2002) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   5  (bases 1 to 143642)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-FEB-2002) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On Jan 4, 2002 this sequence version replaced gi:17647085.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
                Contact: sapiens@watson.wustl.edu
            -------------- Summary Statistics
                Center project name: H_NH0563A13
                Drafting Center: WIBR
            --------------.
```

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Boo

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

□ 1: AC019105. Homo sapiens BAC ...[gi:13677157] Links

```
LOCUS       AC019105               170491 bp    DNA     linear   PRI 07-NOV-2001
DEFINITION  Homo sapiens BAC clone RP11-475A8 from 2, complete sequence.
ACCESSION   AC019105
VERSION     AC019105.7  GI:13677157
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 170491)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
  MEDLINE   99063792
   PUBMED   9847074
REFERENCE   2  (bases 1 to 170491)
  AUTHORS   Belter,E., Doebber,A., Abbott,A. and Ahluwalia,R.
  TITLE     The sequence of Homo sapiens BAC clone RP11-475A8
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 170491)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-DEC-1999) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 170491)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-APR-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   5  (bases 1 to 170491)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-APR-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   6  (bases 1 to 170491)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-NOV-2001) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On Apr 19, 2001 this sequence version replaced gi:10048052.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
```

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Boo

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

□ 1: AC019159. Homo sapiens BAC ...[gi:13677116] Links

```
LOCUS       AC019159               163085 bp    DNA     linear   PRI 07-NOV-2001
DEFINITION  Homo sapiens BAC clone RP11-56O18 from 2, complete sequence.
ACCESSION   AC019159
VERSION     AC019159.8  GI:13677116
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 163085)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
  MEDLINE   99063792
   PUBMED   9847074
REFERENCE   2  (bases 1 to 163085)
  AUTHORS   Goyea,E., Cotton,M., Spalding,L. and Lehnert,L.
  TITLE     The sequence of Homo sapiens BAC clone RP11-56O18
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 163085)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (30-DEC-1999) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 163085)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-APR-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   5  (bases 1 to 163085)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-APR-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   6  (bases 1 to 163085)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-NOV-2001) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On Apr 19, 2001 this sequence version replaced gi:11276269.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
```

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Boo

Search Nucleotide for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

□ 1: AC104648. Homo sapiens BAC ...[gi:18042341] Links

```
LOCUS       AC104648            112084 bp    DNA     linear   PRI 21-FEB-2002
DEFINITION  Homo sapiens BAC clone RP11-45D4 from 2, complete sequence.
ACCESSION   AC104648 AC015602
VERSION     AC104648.2  GI:18042341
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 112084)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
  MEDLINE   99063792
   PUBMED   9847074
REFERENCE   2  (bases 1 to 112084)
  AUTHORS   Bielicki,L. and Abbott,A.
  TITLE     The sequence of Homo sapiens BAC clone RP11-45D4
  JOURNAL   Unpublished (2001)
REFERENCE   3  (bases 1 to 112084)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-DEC-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 112084)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (03-JAN-2002) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   5  (bases 1 to 112084)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-FEB-2002) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On Jan 3, 2002 this sequence version replaced gi:17921241.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
                Contact: sapiens@watson.wustl.edu
            -------------- Summary Statistics
                Center project name: H_NH0045D04
                Drafting Center: WIBR
            --------------.
```

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Boo

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

□ 1: AC074362. Homo sapiens BAC ...[gi:14140337] Links

```
LOCUS       AC074362               149705 bp    DNA     linear   PRI 07-NOV-2001
DEFINITION  Homo sapiens BAC clone RP11-1C10 from 2, complete sequence.
ACCESSION   AC074362
VERSION     AC074362.5  GI:14140337
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 149705)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
  MEDLINE   99063792
   PUBMED   9847074
REFERENCE   2  (bases 1 to 149705)
  AUTHORS   Belter,E., Abbott,A. and Despot,J.
  TITLE     The sequence of Homo sapiens BAC clone RP11-1C10
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 149705)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-JUL-2000) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 149705)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAY-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   5  (bases 1 to 149705)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (07-NOV-2001) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On May 17, 2001 this sequence version replaced gi:13518223.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
                Contact: sapiens@watson.wustl.edu
            -------------- Summary Statistics
                Center project name: H_NH0001C10
            --------------.

            NOTICE: This sequence may not represent the entire insert of this
```

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Boo

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Show: 20 Send to File Get Subsequence

□ 1: AC079154. Homo sapiens BAC ...[gi:15778757] Links

```
LOCUS       AC079154              175387 bp    DNA     linear   PRI 09-JAN-2002
DEFINITION  Homo sapiens BAC clone RP11-314E14 from 2, complete sequence.
ACCESSION   AC079154
VERSION     AC079154.5  GI:15778757
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 175387)
  AUTHORS   Sulston,J.E. and Waterston,R.
  TITLE     Toward a complete human genome sequence
  JOURNAL   Genome Res. 8 (11), 1097-1108 (1998)
  MEDLINE   99063792
   PUBMED   9847074
REFERENCE   2  (bases 1 to 175387)
  AUTHORS   McLellan,M., Cotton,M. and Doebber,A.
  TITLE     The sequence of Homo sapiens BAC clone RP11-314E14
  JOURNAL   Unpublished (2001)
REFERENCE   3  (bases 1 to 175387)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-AUG-2000) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   4  (bases 1 to 175387)
  AUTHORS   Waterston,R.H.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-SEP-2001) Genome Sequencing Center, Washington
            University School of Medicine, 4444 Forest Park Parkway, St. Louis,
            MO 63108, USA
REFERENCE   5  (bases 1 to 175387)
  AUTHORS   Waterston,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (09-JAN-2002) Department of Genetics, Washington
            University, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA
COMMENT     On Sep 26, 2001 this sequence version replaced gi:13654392.
            -------------- Genome Center
                Center: Washington University Genome Sequencing Center
                Center code: WUGSC
                Web site: http://genome.wustl.edu/gsc
                Contact: sapiens@watson.wustl.edu
            -------------- Summary Statistics
                Center project name: H_NH0314E14
            --------------.

            NOTICE:  This sequence may not represent the entire insert of this
```

FF

US005817479A

United States Patent [19]

Au-Young et al.

[11] **Patent Number: 5,817,479**

[45] **Date of Patent: Oct. 6, 1998**

[54] **HUMAN KINASE HOMOLOGS**

[75] Inventors: **Janice Au-Young**, Berkeley; **Olga Bandman**; **Phillip R. Hawkins**, both of Mountain View; **Craig G. Wilde**, Sunnyvale, all of Calif.

[73] Assignee: **Incyte Pharmaceuticals, Inc.**, Palo Alto, Calif.

[21] Appl. No.: **700,575**

[22] Filed: **Aug. 7, 1996**

[51] **Int. Cl.^{6}** **C12P 21/06; C12N 15/64**

[52] **U.S. Cl.** **435/69.1; 435/91.4; 435/320.1; 435/325; 435/252.1; 536/23.2; 536/23.5**

[58] **Field of Search** 536/23.1, 23.2, 536/23.5; 435/91.4, 325, 320.1, 69.1, 252.1

[56] **References Cited**

PUBLICATIONS

Taniguchi, "Cytokine Signaling Through Nonreceptor Protein Tyrosine Kinases," *Science*, 268:251–55 (14 Apr. 1995).

Egan et al., "The pathway to signal achievement," *Nature*, 365:781–783.

Derijard et al., "Independent Human MAP Kinase Signal Transduction Pathways Defined by MEK and MKK Isoforms," *Science*, 267:682–686 (3 Feb. 1995).

R. Davis, "MAPKs: new JNK expands the group," *TIBS*, 19:470–473 (1994).

Han et al., "A MAP Kinase Targeted by Endotoxin and Hyperosmolarity in Mammalian Cells," *Science*, 265:808–811 (1994).

Levitzki et al., "Tyrosine Kinase Inhibition: An Approach to Drug Development," *Science*, 267:1782–1788 (Mar. 24, 1995).

Stroberg, "Functional expression of receptors in microorganisms," *Trends in Pharmacol.*, 13(3)95–98.

Hanes et al. Gen Bank J. Mol. Biol. 244: 665–672, 1994.

Hanes et al. (1994) J. Mol. Biol. 244, 665–672.

Tamagnone et al. (1994) Oncogene 9(12), 3683–3688.

Bennett et al. (1994) J. Biol. Chem. 269(19), 14211–14218.

Primary Examiner—Frank C. Eisenschenk
Assistant Examiner—Patrick J. Nolan
Attorney, Agent, or Firm—Lucy J. Billings; Sheela Mohan-Peterson; Incyte Pharmaceuticals, Inc.

[57] **ABSTRACT**

The present invention provides polynucleotides (kin) which identify and encode novel protein kinases (KIN) expressed in various human cells and tissues. The present invention also provides for antisense sequences and oligonucleotides designed from the nucleotide sequences or their complements. The invention further provides genetically engineered expression vectors and host cells for the production of purified KIN peptides, antibodies capable of binding KIN, and inhibitors specifically bind KIN. The invention specifically provides for diagnostic kits and assays which identify a disorder or disease with altered kinase expression and allow monitoring of patients during drug therapy. These assays utilize oligonucleotides or antibodies produced using the kin polynucleotides.

4 Claims, No Drawings

HUMAN KINASE HOMOLOGS

FIELD OF THE INVENTION

The present invention is in the field of molecular biology; more particularly, the present invention describes nucleic acid sequences for novel human kinase homologs.

BACKGROUND OF THE INVENTION

Kinases regulate many different cell proliferation, differentiation, and signalling processes by adding phosphate groups to proteins. Uncontrolled signalling has been implicated in inflammation, oncogenesis, arteriosclerosis, and psoriasis. Reversible protein phosphorylation is the main strategy for controlling activities of eukaryotic cells. It is estimated that more than 1000 of the 10,000 proteins active in a typical mammalian cell are phosphorylated. The high energy phosphate which drives activation is generally transferred from adenosine triphosphate molecules (ATP) to a particular protein by protein kinases and removed from that protein by protein phosphatases.

Phosphorylation occurs in response to extracellular signals (hormones, neurotransmitters, growth and differentiation factors, etc), cell cycle checkpoints, and environmental or nutritional stresses and is roughly analogous to the turning on a molecular switch. When the switch goes on, the appropriate protein kinase activates a metabolic enzyme, regulatory protein, receptor, cytoskeletal protein, ion channel or pump, or transcription factor.

The kinases comprise the largest known protein family, a superfamily of enzymes with widely varied functions and specificities. They are usually named after their substrate, their regulatory molecules, after some aspect of a mutant phenotype or arbitrarily. Almost all kinases contain a similar 250–300 amino acid catalytic domain. The N-terminal domain, which contains subdomains I–IV, generally folds into a two-lobed structure and binds and orients the ATP (or GTP) donor molecule. The larger C terminal lobe, which contains subdomains VIA–XI, binds the protein substrate and carries out the transfer of the gamma phosphate from ATP to the hydroxyl group of a serine, threonine, or tyrosine residue. Subdomain V spans the two lobes.

The kinases may be categorized into families by the different amino acid sequences (generally between 5 and 100 residues) located on either side of, or inserted into loops of, the kinase domain. These added amino acid sequences allow the regulation of each kinase as it recognizes and interacts with its target protein. The primary structure of the kinase domains is conserved and can be further subdivided into 12 subdomains. The following residues are relatively (~95%) invariant: G_{50} and G_{52} in subdomain I, K_{72} in subdomain II, G_{91} in subdomain III, E_{208} in subdomain VIII, D_{220} and G_{225} in subdomain IX, and the motifs or patterns of amino acids in subdomains VIB, VIII and IX (Hardie G. and Hanks S. (1995) *The Protein Kinase Facts Books*, I and II, Academic Press, San Diego, Calif.).

The cyclin dependent protein kinase (cdk) family includes proteins which are turned on and off as the cell proceeds through the cell cycle. A cdk is active as a kinase only when it is bound to a cyclin. Cdk activation simultaneously requires both the addition of a high energy phosphate to a threonine residue by a kinase and the removal of a covalently-bound phosphate from a specific tyrosine residue by a phosphatase. The concentration of some cyclins rises gradually through a particular part of the cell cycle until their targeted proteolysis ends the coordinated interaction among the cyclin, kinase, and phosphatase molecules.

The second-messenger dependent protein kinases primarily mediate the effects of second messengers such as cyclic AMP (cAMP) cyclic GMP, inositol triphosphate, phosphatidylinositol, 3,4,5-triphosphate, cyclic ADPribose, arachidonic acid and diacylglycerol. For purposes of example, the structure and function of cyclic AMP-dependent protein kinase (A-kinase) will be described. Mammalian cells generally contain at least two forms of A-kinase; type 1 which is cytosolic, and type 2 which is bound to plasma membrane, nuclear membrane or microtubules. In its inactive state, A-kinase consists of a complex of two catalytic subunits and two regulatory subunits. When each regulatory subunit has bound two molecules of cAMP, the catalytic subunit is activated and can transfer a high energy phosphate from ATP to the serine or threonine of a substrate protein. Substrate proteins are usually marked by the presence of two or more basic amino acids on their amino terminal sides. A-kinase is important in metabolism of glycogen, for inactivation of phosphatase inhibitor protein, in transcription of genes which contain a regulatory region called the cAMP response element (CRE), and in regulation of the ion channels of olfactory neurons.

Protein kinase C (PKC) is a water-soluble, Ca^{++}-dependent kinase, commonly found in brain tissue, which moves to the plasma membrane in the presence of Ca^{++} ions. Approximately half of the known isoforms of PKC are activated initially by diacylglycerol and phosphatidylserine. Prolonged activation of PKC depends on continued production of diacyglycerol molecules which are formed when phospholipases cleave phosphatidylcholine. In nerve cells, PKC phosphorylates ion channels and alters the excitability of the cell membrane. In other cells, activation of PKC increases gene transcription either by triggering a protein kinase cascade which activates a regulatory element (much like CRE above) or by phosphorylating and deactivating an inhibitor of the regulatory protein.

Ca^{++}/calmodulin-dependent protein kinases (CaM-kinases) mediate most of the actions of Ca^{++} in human cells. The CaM-kinases include enzymes with narrow substrate specificity such as myosin light chain kinase which activates smooth muscle contraction and phosphorylase kinase which activates glycogen breakdown and the multifunctional enzyme, CaM-kinase II which is found in all cells. Phosphorylase kinase has four subunits: γ is the catalytic moiety and α, β and □δ are regulatory. Since subunits α and β are phosphorylated by A-kinase and subunit □δ is Ca^{++}/calmodulin, glycogen breakdown can be activated by either cAMP or Ca^{++}.

CaM-kinase II is particularly enriched in catecholamine synapses. In those neurons, Ca^{++} influx stimulates both the release of dopamine, noradrenaline or adrenaline and also their resynthesis through the activation of CaM-kinase II. Although the main role of CaM-kinase II is phosphorylation of tyrosine hydroxylase, the rate-limiting enzyme of catecholamine synthesis, CaM-kinase II also autophosphorylates and remains active until phosphotases overwhelm it.

Transmembrane protein-tyrosine kinases are receptors for most growth factors. The first characterized receptor for epidermal growth factor (EGF) is a single pass transmembrane protein of about 1200 amino acids with an extracellular glycosylated portion that interacts with the 53 amino acid EGF molecule. Binding activates the transfer of a phosphate group from ATP to selected tyrosine side chains of the receptor and other specific proteins. Other protein receptors with similar structure include the following growth and differentiation factors (GF)—platelet derived GF, fibroblast GF, hepatocyte GF, insulin and insulin-like GFs, nerve

GF, vascular endothelial GF, macrophage colony stimulating factor, etc. Each protein phosphorylates itself by receptor dimerization to initiate the intracellular signalling cascade.

Many protein-tyrosine kinases lack transmembrane regions and form a complex with the intercellular regions of other cell surface receptors. The best known NR-PTKs are the Src kinase family (Src, Yes, Fgr, Fyn, Lck, Lyn, Hck, Blk, etc) and the Janus kinase family (Jak1, Jak2, Jak3, Tyk2, etc). The Src PTKs are located on the cytoplasmic side of the plasma membrane and are characterized by Src homology regions 2 and 3 (SH2 and SH3). Src PTKs recognize short peptide motifs bearing phosphotyrosine or proline residues, respectively, and mediate protein-protein interactions that regulate a whole range of intracellular signalling molecules. Janus PTKs contain PTK or PTK-like domains and interact with growth hormone, prolactin, and some of the same cytokine receptors as Src PTKs. The cytokine receptors are unique both in their ability to recruit multiple PTKs and in the diversity of their intracellular domains which allow flexibility in their responses within different cell types (Taniguchi T. (1995) Science 268:251–55). Src and Jak kinases were first identified as the products of mutant oncogenes in cancer cells where their activation was no longer subject to normal cellular controls.

Extracellular signalling proteins such as transforming growth factor-β (TGF-β), activins, bone morphogenetic protein, and related members of the TGF-β superfamily interact with receptor serine/threonine kinases. Like EGF above, these receptor kinases have a single pass transmembrane domain with a serine/threonine kinase residue on the cytosolic side of the plasma membrane. The signalling pathways which are activated by binding the extracellular signalling molecules are presently under investigation.

Mitogen-activated protein (MAP) kinases also regulate intracellular signalling pathways. They mediate signal transduction from cell surface to nuclei via phosphorylation cascades. Several subgroups have been identified, and each manifests different substrate specificities and responds to distinct extracellular stimuli (Egan S. E. and Weinberg R. A. (1993) Nature 365:781–783).

MAP kinase signalling pathways are present in mammalian cells as well as in yeast. The extracellular stimuli which activate mammalian pathways include epidermal growth factor (EGF), ultraviolet light, hyperosmolar medium, heat shock, endotoxic lipopolysaccharide (LPS), and pro-inflammatory cytokines such as tumor necrosis factor (TNF) and interleukin-1 (IL-1). In *Saccharomyces cerevisiae*, exposure to mating pheromone or hyperosmolar environments activate the various MAP kinase signalling pathways.

Mammalian cells have at least three subgroups of MAP kinases (Derijard B. et al (1995) Science 267:682–5), each distinguished by a tripeptide motif. They are extracellular signal-regulated protein kinases (ERK) characterized by Thr-Glu-Tyr; c-Jun amino-terminal kinases (JNK) characterized by Thr-Pro-Tyr; and p38 kinase characterized by Thr-Gly-Tyr. Each subgroup is activated by dual phosphorylation of threonine and tyrosine residues by MAP kinase kinases located upstream of the phosphorylation cascade. Activated MAP kinases, in turn, phosphorylate downstream effectors ultimately leading to intracellular changes.

The ERK signal transduction pathway is activated via tyrosine kinase receptors on the plasmalemma. When growth factors bind to tyrosine, they bind to noncatalytic, Src homology (SH) adaptor proteins (SH2-SH3-SH2) and a guanine nucleotide releasing protein (GNRP). GNRP reduces GTP and activates Ras proteins, members of the large family of guanine nucleotide binding proteins (G-proteins). Activated Ras proteins bind to a protein kinase C-Raf-1 and activate the Raf-1 proteins. The activated Raf-1 kinase subsequently phosphorylates MAP kinase kinase (MKK) which, in turn, activate ERKs.

ERKs are proline-directed protein kinases which phosphorylate Ser/Thr-Pro motifs. In fact, cytoplasmic phospholipase A2 (cPLA2) and transcription factor Elk-1 are substrates of ERKs. The ERKs phosphorylate Ser_{505} of cPLA2 thereby increasing its enzymatic activity and resulting in release of arachidonic acid and the formation of lysophospholipids from membrane phospholipids. Likewise, phosphorylation of the transcription factor Elk-1 by ERK ultimately increases transcriptional activity.

JNK is distantly related to the ERK and is similarly activated by dual phosphorylation of Thr and Tyr and by MKK4 (Davis R (1994) TIBS 19:470–473). The JNK signal transduction pathway is also initiated by ultraviolet light, osmotic stress, and the pro-inflammatory cytokines, TNF and IL-1. Phosphorylation of Ser_{63} and Ser_{73} in the NH_2-terminal domain of the transcription factor c-Jun increases transcriptional activity.

p38 is a 41 kD protein containing 360-amino acids. Its dual phosphorylation is activated by the MKK3 and MKK4, heat shock, hyperosmolar medium, IL-1 or LPS endotoxin (Han J. et al (1994) Science 265:808–811). Sepsis produced by LPS is characterized by fever, chills, tachypnea, and tachycardia, and severe cases may result in septic shock which includes hypotension and multiple organ failure.

Cells respond to LPS as a stress signal because it alters normal cellular processes and induces the release of systemic mediators such as TNF. CD14 is a glycosylphosphatidyl-inositol-anchored membrane glycoprotein which serves as a LPS receptor on the plasmalemma of monocytic cells. The binding of LPS to CD14 causes rapid protein tyrosine phosphorylation of the 44- and 42-/40-kD isoforms of MAP kinases. Although they bind LPS, these MAP kinase isoforms do not appear to belong to the p38 subgroup.

An detailed understanding of kinase pathways and signal transduction is beginning to reveal some mechanisms for interceding in the progression of inflammatory illnesses and of uncontrolled cell proliferation. The cDNAs, oligonucleotides, peptides and antibodies for the human kinases, which are the subject of this invention and are listed in Table 1, provide a plurality of tools for studying signalling cascades in various cells and tissues and for diagnosing and selecting inhibitors or drugs with the potential to intervene in various disorders or diseases in which altered kinase expression is implicated. The disorders or diseases include, but not limited to, human X-linked agammaglobulinemia, nonspherocytic hemolytic anemia, atherosclerosis, carcinomas (breast, ovary, renal, squamous cell and prostate), diabetes, gliomas, glomerular disease, hepatomegaly, Karposi's sarcoma, lymphoblastic and myelogenous leukemias, myoglobinuria, peptic ulcer disease, psoriasis, pulmonary fibrosis, restenosis, and septic shock due to cholera, *Clostridium difficile, E. coli* and Shigella (Isselbacher K. J. et al (1994) Harrison's Principles of Internal Medicine, McGraw-Hill, New York City; Levitzki A. and A. Gazit (1995) Science 267:1782–88).

SUMMARY OF THE INVENTION

The subject invention provides unique polynucleotides (SEQ ID NOs 1–44) which have been identified as novel human kinases (kin). These partial cDNAs were identified among the polynucleotides which comprise various Incyte cDNA libraries.

The invention comprises polynucleotides which are complementary to the kin sequences (SEQ ID Nos 1–44).

The invention also comprises the use of kin sequences to identify and obtain a full length human kinase cDNAs such as SEQ ID NO 45.

The invention further comprises the use of oligomers from these kin sequences in a kinases kit which can be used to identify a disorder or disease with altered kinase expression and provide a method for monitoring progress of a patient during drug therapy.

Aspects of the invention include use of kin sequences or recombinant nucleic acids derived from them to produce purified peptides. Still further aspects of the invention use these purified peptides to identify antibodies or other molecules with inhibitory activity toward a particular kinase, group of kinases or disease.

In addition, the invention comprises the use of kin specific antibodies in assays to identify a disorder or disease with altered kinase expression and provides a method to monitor the progress of a patient during drug therapy.

DESCRIPTION OF THE FIGURE

FIGS. 1A and 1B display the full length nucleotide sequence for human MAP kinase from stomach tissue (SEQ ID NO 45; Incyte Clone 214915E) and its predicted amino acid sequence.

DETAILED DESCRIPTION OF THE INVENTION

Definitions

As used herein, the abbreviation for kinase in lower case (kin) refers to a gene, cDNA, RNA or nucleic acid sequence while the upper case version (KIN) refers to a protein, polypeptide, peptide, oligopeptide, or amino acid sequence.

An "oligonucleotide" or "oligomer" is a stretch of nucleotide residues which has a sufficient number of bases to be used in a polymerase chain reaction (PCR). These short sequences are based on (or designed from) genomic or cDNA sequences and are used to amplify, confirm, or reveal the presence of an identical, similar or complementary DNA or RNA in a particular cell or tissue. Oligonucleotides or oligomers comprise portions of a DNA sequence having at least about 10 nucleotides and as many as about 50 nucleotides, preferably about 15 to 30 nucleotides. They are chemically synthesized and may be used as probes.

"Probes" are nucleic acid sequences of variable length, preferably between at least about 10 and as many as about 6,000 nucleotides, depending on use. They are used in the detection of identical, similar, or complementary nucleic acid sequences. Longer length probes are usually obtained from a natural or recombinant source, are highly specific and much slower to hybridize than oligomers. They may be single- or double-stranded and carefully designed to have specificity in PCR, hybridization membrane-based, or ELISA-like technologies.

"Reporter" molecules are chemical moieties used for labelling a nucleic or amino acid sequence. They include, but are not limited to, radionuclides, enzymes, fluorescent, chemi-luminescent, or chromogenic agents. Reporter molecules associate with, establish the presence of, and may allow quantification of a particular nucleic or amino acid sequence.

A "portion" or "fragment" of a polynucleotide or nucleic acid comprises all or any part of the nucleotide sequence having fewer nucleotides than about 6 kb, preferably fewer than about 1 kb which can be used as a probe. Such probes may be labelled with reporter molecules using nick translation, Klenow fill-in reaction, PCR or other methods well known in the art. After pretesting to optimize reaction conditions and to eliminate false positives, nucleic acid probes may be used in Southern, northern or in situ hybridizations to determine whether DNA or RNA encoding the protein is present in a biological sample, cell type, tissue, organ or organism.

"Recombinant nucleotide variants" are polynucleotides which encode a protein. They may be synthesized by making use of the "redundancy" in the genetic code. Various codon substitutions, such as the silent changes which produce specific restriction sites or codon usage-specific mutations, may be introduced to optimize cloning into a plasmid or viral vector or expression in a particular prokaryotic or eukaryotic host system, respectively.

"Linkers" are synthesized palindromic nucleotide sequences which create internal restriction endonuclease sites for ease of cloning the genetic material of choice into various vectors. "Polylinkers" are engineered to include multiple restriction enzyme sites and provide for the use of both those enzymes which leave 5' and 3' overhangs such as BamHI, EcoRI, PstI, KpnI and Hind III or which provide a blunt end such as EcoRV, SnaBI and StuI.

"Control elements" or "regulatory sequences" are those nontranslated regions of the gene or DNA such as enhancers, promoters, introns and 3' untranslated regions which interact with cellular proteins to carry out replication, transcription, and translation. They may occur as boundary sequences or even split the gene. They function at the molecular level and along with regulatory genes are very important in development, growth, differentiation and aging processes.

"Chimeric" molecules are polynucleotides or polypeptides which are created by combining one or more of nucleotide sequences of this invention (or their parts) with additional nucleic acid sequence(s). Such combined sequences may be introduced into an appropriate vector and expressed to give rise to a chimeric polypeptide which may be expected to be different from the native molecule in one or more of the following kinase characteristics: cellular location, distribution, ligand-binding affinities, interchain affinities, degradation/turnover rate, signalling, etc.

"Active" is that state which is capable of being useful or of carrying out some role. It specifically refers to those forms, fragments, or domains of an amino acid sequence which display the biologic and/or immunogenic activity characteristic of the naturally occurring kinase.

"Naturally occurring KIN" refers to a polypeptide produced by cells which have not been genetically engineered or which have been genetically engineered to produce the same sequence as that naturally produced. Specifically contemplated are various polypeptides which arise from post-transnational modifications. Such modifications of the polypeptide include but are not limited to acetylation, carboxylation, glycosylation, phosphorylation, lipidation and acylation.

"Derivative" refers to those polypeptides which have been chemically modified by such techniques as ubiquitination, labelling (see above), pegylation (derivatization with polyethylene glycol), and chemical insertion or substitution of amino acids such as ornithine which do not normally occur in human proteins.

"Recombinant polypeptide variant" refers to any polypeptide which differs from naturally occurring KIN by amino acid insertions, deletions and/or substitutions, created using recombinant DNA techniques. Guidance in determining which amino acid residues may be replaced, added or deleted without abolishing characteristics of interest may be found by comparing the sequence of KIN with that of related polypeptides and minimizing the number of amino acid sequence changes made in highly conserved regions.

Amino acid "substitutions" are defined as one for one amino acid replacements. They are conservative in nature when the substituted amino acid has similar structural and/or chemical properties. Examples of conservative replacements are substitution of a leucine with an isoleucine or valine, an aspartate with a glutamate, or a threonine with a serine.

Amino acid "insertions" or "deletions" are changes to or within an amino acid sequence. They typically fall in the range of about 1 to 5 amino acids. The variation allowed in a particular amino acid sequence may be experimentally determined by producing the peptide synthetically or by systematically making insertions, deletions, or substitutions of nucleotides in the kin sequence using recombinant DNA techniques.

A "signal or leader sequence" is a short amino acid sequence which or can be used, when desired, to direct the polypeptide through a membrane of a cell. Such a sequence may be naturally present on the polypeptides of the present invention or provided from heterologous sources by recombinant DNA techniques.

An "oligopeptide" is a short stretch of amino acid residues and may be expressed from an oligonucleotide. It may be functionally equivalent to and either the same length as or considerably shorter than a "fragment ", "portion ", or "segment" of a polypeptide. Such sequences comprise a stretch of amino acid residues of at least about 5 amino acids and often about 17 or more amino acids, typically at least about 9 to 13 amino acids, and of sufficient length to display biologic and/or immunogenic activity.

An "inhibitor" is a substance which retards or prevents a chemical or physiological reaction or response. Common inhibitors include but are not limited to antisense molecules, antibodies, antagonists and their derivatives.

A "standard" is a quantitative or qualitative measurement for comparison. Preferably, it is based on a statistically appropriate number of samples and is created to use as a basis of comparison when performing diagnostic assays, running clinical trials, or following patient treatment profiles. The samples of a particular standard may be normal or similarly abnormal.

"Animal" as used herein may be defined to include human, domestic (cats, dogs, etc), agricultural (cows, horses, sheep, goats, chicken, fish, etc) or test species (frogs, mice, rats, rabbits, simians, etc).

"Disorders or diseases" in which altered kinase activity have been implicated specifically include, but are not limited to, human X-linked agammaglobulinemia, nonspherocytic hemolytic anemia, atherosclerosis, carcinomas (breast, ovary, renal, squamous cell and prostate), diabetes, gliomas, glomerular disease, hepatomegaly, Karposi's sarcoma, lymphoblastic and myelogenous leukemias, myoglobinuria, peptic ulcer disease, psoriasis, pulmonary fibrosis, restenosis, and septic shock due to cholera, *Clostridium difficile, E. coli* and Shigella.

Since the list of technical and scientific terms cannot be all encompassing, any undefined terms shall be construed to have the same meaning as is commonly understood by one of skill in the art to which this invention belongs. Furthermore, the singular forms "a", "an" and "the" include plural referents unless the context clearly dictates otherwise. For example, reference to a "restriction enzyme" or a "high fidelity enzyme" may include mixtures of such enzymes and any other enzymes fitting the stated criteria, or reference to the method includes reference to one or more methods for obtaining cDNA sequences which will be known to those skilled in the art or will become known to them upon reading this specification.

Before the present sequences, variants, formulations and methods for making and using the invention are described, it is to be understood that the invention is not to be limited only to the particular sequences, variants, formulations or methods described. The sequences, variants, formulations and methodologies may vary, and the terminology used herein is for the purpose of describing particular embodiments. The terminology and definitions are not intended to be limiting since the scope of protection will ultimately depend upon the claims.

DESCRIPTION OF THE INVENTION

The present invention provides for purified partial protein kinase cDNAs which were expressed in various human tissues and isolated therefrom. These sequences were identified by their similarity to published or known open reading frames or untranslated control regions. Since protein kinases are associated with basic cellular processes such as cell proliferation, differentiation and cell signalling, these nucleotide sequences are useful in the characterization of and delineation of normal and abnormal processes. Kinase nucleotide sequences are useful in diagnostic assays used to evaluate the role of a specific kinase in normal, diseased, or therapeutically treated cells.

Purified kinase nucleotide sequences have numerous applications in techniques known to those skilled in the art of molecular biology. These techniques include their use as hybridization probes, for chromosome and gene mapping, in PCR technologies, in the production of sense or antisense nucleic acids, in screening for new therapeutic molecules, etc. These examples are well known and are not intended to be limiting. Furthermore, the nucleotide sequences disclosed herein may be used in molecular biology techniques that have not yet been developed, provided the new techniques rely on properties of nucleotide sequences that are currently known, including but not limited to such properties as the triplet genetic code and specific base pair interactions.

As a result of the degeneracy of the genetic code, a multitude of kinase-encoding nucleotide sequences may be produced and some of these will bear only minimal homology to the endogenous sequence of any known and naturally occurring kinase. This invention has specifically contemplated each and every possible variation of nucleotide sequence that could be made by selecting combinations based on possible codon choices. These combinations are made in accordance with the standard triplet genetic code as applied to the nucleotide sequence of naturally occurring kinases, and all such variations are to be considered as being specifically disclosed.

Although the kinase nucleotide sequences and their derivatives or variants are preferably capable of identifying the nucleotide sequence of the naturally occurring kinase under optimized conditions, it may be advantageous to produce kinase-encoding nucleotide sequences possessing a substantially different codon usage. Codons can be selected to increase the rate at which expression of the peptide occurs in a particular prokaryotic or eukaryotic expression host in accordance with the frequency with which particular codons are utilized by the host. Other reasons for substantially altering the nucleotide sequence encoding the kinase without altering the encoded amino acid sequence include the production of RNA transcripts having more desirable properties, such as a longer half-life, than transcripts produced from the naturally occurring sequence.

Nucleotide sequences encoding a kinase may be joined to a variety of other nucleotide sequences by means of well established recombinant DNA techniques (Sambrook J. et al (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.; or Ausubel F. M. et al (1989) Current Protocols in Molecular Biology, John Wiley & Sons, New York City). Useful sequences for joining to the kinase include an assortment of cloning vectors such as plasmids, cosmids, lambda phage derivatives, phagemids, and the like. Vectors of interest include vectors for replication, expression, probe generation, sequencing, and the like. In general, vectors of interest may contain an origin of replication functional in at least one organism, convenient restriction endonuclease sensitive sites, and selectable markers for one or more host cell systems.

PCR as described in U.S. Pat. Nos. 4,683,195; 4,800,195; and 4,965,188 provides additional uses for oligonucleotides based upon the kinase nucleotide sequence. Such oligomers are generally chemically synthesized, but they may be of recombinant origin or a mixture of both. Oligomers generally comprise two nucleotide sequences, one with sense orientation (5'→3') and one with antisense (3' to 5') employed under optimized conditions for identification of a specific gene or diagnostic use. The same two oligomers, nested sets of oligomers, or even a degenerate pool of oligomers may be employed under less stringent conditions for identification and/or quantitation of closely related DNA or RNA sequences.

Full length genes may be cloned utilizing partial nucleotide sequence and various methods known in the art. Gobinda et al (1993; PCR Methods Applic 2:318–22) disclose "restriction-site PCR" as a direct method which uses universal primers to retrieve unknown sequence adjacent to a known locus. First, genomic DNA is amplified in the presence of primer to linker and a primer specific to the known region. The amplified sequences are subjected to a second round of PCR with the same linker primer and another specific primer internal to the first one. Products of each round of PCR are transcribed with an appropriate RNA polymerase and sequenced using reverse transcriptase. Gobinda et al present data concerning Factor IX for which they identified a conserved stretch of 20 nucleotides in the 3' noncoding region of the gene.

Inverse PCR is the first method to report successful acquisition of unknown sequences starting with primers based on a known region (Triglia T. et al (1988) Nucleic Acids Res 16:8186). The method uses several restriction enzymes to generate a suitable fragment in the known region of a gene. The fragment is then circularized by intramolecular ligation and used as a PCR template. Divergent primers are designed from the known region. The multiple rounds of restriction enzyme digestions and ligations that are necessary prior to PCR make the procedure slow and expensive (Gobinda et al, supra).

Capture PCR (Lagerstrom M. et al (1991) PCR Methods Applic 1:111–19) is a method for PCR amplification of DNA fragments adjacent to a known sequence in human and YAC DNA. As noted by Gobinda et al (supra), capture PCR also requires multiple restriction enzyme digestions and ligations to place an engineered double-stranded sequence into an unknown portion of the DNA molecule before PCR. Although the restriction and ligation reactions are carried out simultaneously, the requirements for extension, immobilization and two rounds of PCR and purification prior to sequencing render the method cumbersome and time consuming.

Parker J. D. et al (1991; Nucleic Acids Res 19:3055–60), teach walking PCR, a method for targeted gene walking which permits retrieval of unknown sequence. PromoterFinder™ is a new kit available from Clontech (Palo Alto, Calif.) which uses PCR and primers derived from p53 to walk in genomic DNA. Nested primers and special PromoterFinder libraries are used to detect upstream sequences such as promoters and regulatory elements. This process avoids the need to screen libraries and is useful in finding intron/exon junctions.

Another new PCR method, "Improved Method for Obtaining Full Length cDNA Sequences" by Guegler et al, patent application Ser. No 08/487,112, filed Jun. 7, 1995 and hereby incorporated by reference, employs XL-PCR (Perkin-Elmer, Foster City, Calif.) to amplify and extend partial nucleotide sequence into longer pieces of DNA. This method was developed to allow a single researcher to process multiple genes (up to 20 or more) at one time and to obtain an extended (possibly full-length) sequence within 6–10 days. This new method replaces methods which use labelled probes to screen plasmid libraries and allow one researcher to process only about 3–5 genes in 14–40 days.

In the first step, which can be performed in about two days, any two of a plurality of primers are designed and synthesized based on a known partial sequence. In step 2, which takes about six to eight hours, the sequence is extended by PCR amplification of a selected library. Steps 3 and 4, which take about one day, are purification of the amplified cDNA and its ligation into an appropriate vector. Step 5, which takes about one day, involves transforming and growing up host bacteria. In step 6, which takes approximately five hours, PCR is used to screen bacterial clones for extended sequence. The final steps, which take about one day, involve the preparation and sequencing of selected clones.

If the full length cDNA has not been obtained, the entire procedure is repeated using either the original library or some other preferred library. The preferred library may be one that has been size-selected to include only larger cDNAs or may consist of single or combined commercially available libraries, eg. lung, liver, heart and brain from Gibco/BRL (Gaithersburg, Md.). The cDNA library may have been prepared with oligo (dT) or random priming. Random primed libraries are preferred in that they will contain more sequences which contain 5' ends of genes. A randomly primed library may be particularly useful if an oligo (dT) library does not yield a complete gene. It must be noted that the larger and more complex the protein, the less likely it is that the complete gene will be found in a single plasmid.

A new method for analyzing either the size or the nucleotide sequence of PCR products is capillary electrophoresis. Systems for rapid sequencing are available from Perkin Elmer (Foster, City Calif.), Beckman Instruments (Fullerton, Calif.), and other companies. Capillary sequencing employs flowable polymers for electrophoretic separation, four different fluorescent dyes (one for each nucleotide) which are laser activated, and detection of the emitted wavelengths by a charge coupled devise camera. Output/light intensity is converted to electrical signal using appropriate software (eg. Genotyper™ and Sequence Navigators™ from Perkin Elmer) and the entire process from loading of samples to computer analysis and electronic data display is computer controlled. Capillary electrophoresis provides greater resolution and is many times faster than standard gel based procedures. It is particularly suited to the sequencing of small pieces of DNA which might be present in limited amounts in a particular sample. The reproducible sequencing of up to 350 bp of M13 phage DNA in 30 min has been reported (Ruiz-Martinez M. C. et al (1993) Anal Chem 65:2851–8).

Another aspect of the subject invention is to provide for kinase hybridization probes which are capable of hybridizing with naturally occurring nucleotide sequences encoding kinases. The stringency of the hybridization conditions will determine whether the probe identifies only the native nucleotide sequence of that specific kinase or sequences of closely related molecules. If degenerate kinase nucleotide sequences of the subject invention are used for the detection of related kinase encoding sequences, they should preferably contain at least 50% of the nucleotides of the sequences presented herein. Hybridization probes of the subject invention may be derived from the nucleotide sequences of the SEQ ID NOs 1–44, or from surrounding or included genomic sequences comprising untranslated regions such as promoters, enhancers and introns. Such hybridization probes may be labelled with appropriate reporter molecules. Means for producing specific hybridization probes for kinases include oligolabelling, nick translation, end-labelling or PCR amplification using a labelled nucleotide. Alternatively, the cDNA sequence may be cloned into a vector for the production of mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by addition of an appropriate RNA polymerase such as T7, T3 or SP6 and labelled nucleotides. A number of companies (such as Pharmacia Biotech, Piscataway, N.J.; Promega, Madison, Wis.; US Biochemical Corp, Cleveland, Ohio; etc.) supply commercial kits and protocols for these procedures.

It is also possible to produce a DNA sequence, or portions thereof, entirely by synthetic chemistry. Sometimes the source of information for producing this sequence comes from the known homologous sequence from closely related organisms. After synthesis, the nucleic acid sequence can be used alone or joined with a preexisting sequence and inserted into one of the many available DNA vectors and their respective host cells using techniques well known in the art. Moreover, synthetic chemistry may be used to introduce specific mutations into the nucleotide sequence. Alternatively, a portion of sequence in which a mutation is desired can be synthesized and recombined with a portion of an existing genomic or recombinant sequence.

The kinase nucleotide sequences can be used individually, or in panels, in a diagnostic test or assay to detect disorder or disease processes associated with abnormal levels of kinase expression. The nucleotide sequence is added to a sample (fluid, cell or tissue) from a patient under hybridizing conditions. After an incubation period, the sample is washed with a compatible fluid which optionally contains a reporter molecule which will bind the specific nucleotide. After the compatible fluid is rinsed off, the reporter molecule is quantitated and compared with a standard for that fluid, cell or tissue. If kinase expression is significantly different from the standard, the assay indicates the presence of disorder or disease. The form of such qualitative or quantitative methods may include northern analysis, dot blot or other membrane based technologies, dip stick, pin or chip technologies, PCR, ELISAs or other multiple sample format technologies.

This same assay, combining a sample with the nucleotide sequence, is applicable in evaluating the efficacy of a particular therapeutic treatment regime. It may be used in animal studies, in clinical trials, or in monitoring the treatment of an individual patient. First, standard expression must be established for use as a basis of comparison. Second, samples from the animals or patients affected by the disorder or disease are combined with the nucleotide sequence to evaluate the deviation from the standard or normal profile. Third, an existing therapeutic agent is administered, and a treatment profile is generated. The assay is evaluated to determine whether the profile progresses toward or returns to the standard pattern. Successive treatment profiles may be used to show the efficacy of treatment over a period of several days or several months.

The nucleotide sequence for any particular kinase (SEQ ID NOs 1–45) can also be used to generate probes for mapping the native genomic sequence. The sequence may be mapped to a particular chromosome or to a specific region of the chromosome using well known techniques. These include in situ hybridization to chromosomal spreads (Verma et al (1988) Human Chromosomes: A Manual of Basic Techniques, Pergamon Press, New York City), flow-sorted chromosomal preparations, or artificial chromosome constructions such as yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs), bacterial P1 constructions or single chromosome cDNA libraries.

In situ hybridization of chromosomal preparations and physical mapping techniques such as linkage analysis using established chromosomal markers are invaluable in extending genetic maps. Examples of genetic maps can be found in the 1994 Genome Issue of Science (265:1981f). Often the placement of a gene on the chromosome of another mammalian species may reveal associated markers even if the number or arm of a particular human chromosome is not known. New partial nucleotide sequences can be assigned to chromosomal arms, or parts thereof, by physical mapping. This provides valuable information to investigators searching for disease genes using positional cloning or other gene discovery techniques. Once a disease or syndrome, such as ataxia telangiectasia (AT), has been crudely localized by genetic linkage to a particular genomic region, for example, AT to 11q22-23 (Gatti et al (1988) Nature 336:577–580), any sequences mapping to that area may represent genes for further investigation. The nucleotide sequences of the subject invention may also be used to detect differences in the chromosomal location of nucleotide sequences due to translocation, inversion, etc. between normal and carrier or affected individuals.

The partial nucleotide sequence encoding a particular kinase may be used to produce an amino acid sequence using well known methods of recombinant DNA technology. Goeddel (1990, Gene Expression Technology, Methods and Enzymology, Vol 185, Academic Press, San Diego, Calif.) is one among many publications which teach expression of an isolated, purified nucleotide sequence. The amino acid or peptide may be expressed in a variety of host cells, either prokaryotic or eukaryotic. Host cells may be from the same species from which the nucleotide sequence was derived or from a different species. Advantages of producing an amino acid sequence or peptide by recombinant DNA technology include obtaining adequate amounts for purification and the availability of simplified purification procedures.

Cells transformed with a kinase nucleotide sequence may be cultured under conditions suitable for the expression and recovery of peptide from cell culture. The peptide produced by a recombinant cell may be secreted or may be contained intracellularly depending on the sequence itself and/or the vector used. In general, it is more convenient to prepare recombinant proteins in secreted form, and this is accomplished by ligating kin to a recombinant nucleotide sequence which directs its movement through a particular prokaryotic or eukaryotic cell membrane. Other recombinant constructions may join kin to nucleotide sequence encoding a polypeptide domain which will facilitate protein purification (Kroll D. J. et al (1993) DNA Cell Biol 12:441–53).

Direct peptide synthesis using solid-phase techniques (Stewart et al (1969) Solid-Phase Peptide Synthesis, WH Freeman Co, San Francisco, Calif.; Merrifield J. (1963) J Am Chem Soc 85:2149–2154) is an alternative to recombinant or chimeric peptide production. Automated synthesis may be achieved, for example, using Applied Biosystems 431A Peptide Synthesizer in accordance with the instructions provided by the manufacturer. Additionally a particular kinase sequence or any part thereof may be mutated during direct synthesis and combined using chemical methods with other kinase sequence(s) or a part thereof. This chimeric nucleotide sequence can also be placed in an appropriate vector and host cell to produce a variant peptide.

Although an amino acid sequence or oligopeptide used for antibody induction does not require biological activity, it must be immunogenic. KIN used to induce specific antibodies may have an amino acid sequence consisting of at least five amino acids and preferably at least 10 amino acids. Short stretches of amino acid sequence may be fused with those of another protein such as keyhole limpet hemocyanin, and the chimeric peptide used for antibody production. Alternatively, the oligopeptide may be of sufficient length to contain an entire domain.

Antibodies specific for KIN may be produced by inoculation of an appropriate animal with an antigenic fragment of the peptide. An antibody is specific for KIN if it is produced against an epitope of the polypeptide and binds to at least part of the natural or recombinant protein. Antibody production includes not only the stimulation of an immune response by injection into animals, but also analogous processes such as the production of synthetic antibodies, the screening of recombinant immunoglobulin libraries for specific-binding molecules (Orlandi R. et al (1989) PNAS 86:3833–3837, or Huse W. D. et al (1989) Science 256:1275–1281), or the in vitro stimulation of lymphocyte populations. Current technology (Winter G. and Milstein C. (1991) Nature 349:293–299) provides for a number of highly specific binding reagents based on the principles of antibody formation. These techniques may be adapted to produce molecules which specifically bind kinase peptides. Antibodies or other appropriate molecules generated against a specific immunogenic peptide fragment or oligopeptide can be used in Western analysis, enzyme-linked immunosorbent assays (ELISA) or similar tests to establish the presence of or to quantitate amounts of kinase active in normal, diseased, or therapeutically treated cells or tissues.

The examples below are provided to illustrate the subject invention. These examples are provided by way of illustration and are not included for the purpose of limiting the invention.

EXAMPLES

I cDNA Library Construction

The kinase sequences of this application (Table 1) were first identified among the sequences comprising various libraries. Technology has advanced considerably since the first cDNA libraries were made. Many small variations in both chemicals and machinery have been instituted over time, and these have improved both the efficiency and safety of the process. Although the cDNAs could be obtained using an older procedure, the procedure presented in this application is exemplary of one currently being used by persons skilled in the art. For the purpose of providing an exemplary method, the tissue preparation, mRNA isolation and cDNA library construction described here is for the rheumatoid synovium library from which the Incyte Clones 191283 and 192268 for ser/thr kinases were obtained.

Rheumatoid synovial tissue was obtained from the hip joint removed from a 68 year old female with erosive, nodular rheumatoid arthritis. The tissue was frozen, ground to powder in a mortar and pestle, and lysed immediately in buffer containing guanidinium isothiocyanate. The lysate was centrifuged over a CsCl cushion (18 hrs at 25,000 rpm using a Beckman SW28 rotor and ultracentrifuge; Beckman Instruments, Palo Alto, Calif.), ethanol precipitated, resuspended in water and DNase treated for 15 min at 37° C. The RNA was extracted with phenol chloroform and precipitated with ethanol. Polyadenylated messages were isolated using Qiagen Oligotex (QIAGEN Inc, Chatsworth, Calif.), and a custom cDNA library was constructed by Stratagene (La Jolla, Calif.).

First strand cDNA synthesis was accomplished using an oligo (dT) primer/linker which also contained an XhoI restriction site. Second strand synthesis was performed using a combination of DNA polymerase I, *E. coli* ligase and RNase H, followed by the addition of an EcoRI linker to the blunt ended cDNA. The EcoRI linked, double-stranded cDNA was then digested with XhoI restriction enzyme, extracted with phenol chloroform, and fractionated by size on Sephacryl S400. DNA of the appropriate size was then ligated to dephosphorylated Lambda Zap® arms (Stratagene) and packaged using Gigapack extracts (Stratagene). pBluescript (Stratagene) phagemid DNAs were excised en masse from the library.

In the alternative, DNAs were purified using Miniprep Kits (Catalog #77468; Advanced Genetic Technologies Corporation, Gaithersburg, Md.). These kits provide a 96-well format and enough reagents for 960 purifications. The recommended protocol supplied with each kit has been employed except for the following changes. First, the 96 wells are each filled with only 1 ml of sterile Terrific broth (LIFE TECHNOLGIES™, Gaithersburg, Md.) with carbenicillin at 25 mg/L (2×Carb) and glycerol at 0.4%. After the wells are inoculated, the bacteria are cultured for 24 hours and lysed with 60 μl of lysis buffer. A centrifugation step (2900 rpm for 5 minutes) is performed before the contents of the block are added to the primary filter plate. The optional step of adding isopropanol to TRIS buffer is not routinely performed. After the last step in the protocol, samples are transferred to a Beckman 96-well block for storage.

II Sequencing of cDNA Clones

The cDNA inserts from random isolates of the rheumatoid synovium or other appropriate library were sequenced in part. Methods for DNA sequencing are well known in the art and employ such enzymes as the Klenow fragment of DNA polymerase I, SEQUENASE® (US Biochemical Corp) or Taq polymerase. Methods to extend the DNA from an oligonucleotide primer annealed to the DNA template of interest have been developed for both single- and double-stranded templates. Chain termination reaction products were separated using electrophoresis and detected via their incorporated, labelled precursors. Recent improvements in mechanized reaction preparation, sequencing and analysis have permitted expansion in the number of sequences that can be determined per day. Preferably, the process is automated with machines such as the Hamilton Micro Lab 2200

(Hamilton, Reno, Nev.), Peltier Thermal Cycler (PTC200; MJ Research, Watertown Mass.) and the Applied Biosystems Catalyst 800 and 377 and 373 DNA sequencers.

The quality of any particular cDNA library may be determined by performing a pilot scale analysis of 192 cDNAs and checking for percentages of clones containing vector, lambda or *E. coli* DNA, mitochondrial or repetitive DNA, and clones with exact or homologous matches to public databases. The number of unique sequences—those having no known match in any available database—were recorded.

III Homology Searching of cDNA Clones and Their Deduced Proteins

Each sequence so obtained was compared to sequences in GenBank using a search algorithm developed by Applied Biosystems and incorporated into the INHERIT™ 670 Sequence Analysis System. In this algorithm, Pattern Specification Language (TRW Inc, Los Angeles, Calif.) was used to determine regions of homology. The three parameters that determine how the sequence comparisons run were window size, window offset, and error tolerance. Using a combination of these three parameters, the DNA database was searched for sequences containing regions of homology to the query sequence, and the appropriate sequences were scored with an initial value. Subsequently, these homologous regions were examined using dot matrix homology plots to distinguish regions of homology from chance matches. Smith-Waterman alignments were used to display the results of the homology search.

Peptide and protein sequence homologies were ascertained using the INHERIT™ 670 Sequence Analysis System in a way similar to that used in DNA sequence homologies. Pattern Specification Language and parameter windows were used to search protein databases for sequences containing regions of homology which were scored with an initial value. Dot-matrix homology plots were examined to distinguish regions of significant homology from chance matches.

Alternatively, BLAST, which stands for Basic Local Alignment Search Tool, is used to search for local sequence alignments (Altschul S. F. (1993) J Mol Evol 36:290–300; Altschul, S. F. et al (1990) J Mol Biol 215:403–10). BLAST produces alignments of both nucleotide and amino acid sequences to determine sequence similarity. Because of the local nature of the alignments, BLAST is especially useful in determining exact matches or in identifying homologs. While it is useful for matches which do not contain gaps, it is inappropriate for performing motif-style searching. The fundamental unit of BLAST algorithm output is the High-scoring Segment Pair (HSP).

An HSP consists of two sequence fragments of arbitrary but equal lengths whose alignment is locally maximal and for which the alignmentBLAST approach is to look threshold or cutoff score set by the user. The BLAST approach is to look for HSPs between a query sequence and a database sequence, to evaluate the statistical significance of any matches found, and to report only those matches which satisfy the user-selected threshold of significance. The parameter E establishes the statistically significant threshold for reporting database sequence matches. E is interpreted as the upper bound of the expected frequency of chance occurrence of an HSP (or set of HSPs) within the context of the entire database search. Any database sequence whose match satisfies E is reported in the program output.

All the kinase molecules presented in this application were examined using INHERIT. Although their identification was based on the criteria above, their homology to known kinase molecules and name are subject to change when additional computer analysis against additional or more recent database information is employed. For example, whereas the first two kinases in Table 1 were initially identified as unique Incyte clones, homologous mouse and human kinases are now known. In other cases, additional sequence information has become available and its review against the known databases has precipitated a name change. Occasionally a clone number will also disappear from the LIFESEQ™ database (Incyte Pharmaceuticals Inc, Palo Alto, Calif.). This situation generally arises during the regular review of clones and assembly of contiguous sequences.

IV Extension of cDNAs to Full Length

The kinase sequences presented here can be used to design oligonucleotide primers for the extension of the cDNAs to full length. In fact, the partial map kinase cDNA sequence (SEQ ID NO 38) initially identified in Incyte clone 214915 among the sequences comprising the human stomach cell library was extended to full length as shown in "A Novel Human Map Kinase Homolog" by Hawkins et al. Incyte Docket PF-036P, filed on Jun. 28, 1995, incorporated herein by reference. The coding region of this full length sequence (SEQ ID NO 45; Incyte Clone 214915E) begins at nucleotide 58 and ends at nucleotide 1156.

Primers are designed based on known sequence; one primer is synthesized to initiate extension in the antisense direction (XLR) and the other to extend sequence in the sense direction (XLF). The primers allow the sequence to be extended "outward" generating amplicons containing new, unknown nucleotide sequence for the gene of interest. The primers may be designed using Oligo 4.0 (National Biosciences Inc, Plymouth, Minn.), or another appropriate program, to be 22–30 nucleotides in length, to have a GC content of 50% or more, and to anneal to the target sequence at temperatures about 68°–72° C. Any stretch of nucleotides which would result in hairpin structures and primer-primer dimerizations was avoided.

The stomach cDNA library was used as a template, and XLR=AAG ACA TCC AGG AGC CCA ATG AC and XLF=AGG TGA TCC TCA GCT GGA TGC AC primers were used to extend and amplify the 214915 sequence. By following the instructions for the XL-PCR kit and thoroughly mixing the enzyme and reaction mix, high fidelity amplification is obtained. Beginning with 25 pMol of each primer and the recommended concentrations of all other components of the kit, PCR is performed using the Peltier Thermal Cycler (PTC200; MJ Research, Watertown, Mass.) and the following parameters:

Step 1 94° C. for 60 sec (initial denaturation)
Step 2 94° C. for 15 sec
Step 3 65° C. for 1 min
Step 4 68° C. for 7 min
Step 5 Repeat step 2–4 for 15 additional cycles
Step 6 94° C. for 15 sec
Step 7 65° C. for 1 min
Step 8 68° C. for 7 min+15 sec/cycle
Step 9 Repeat step 6–8 for 11 additional cycles
Step 10 72° C. for 8 min
Step 11 4° C. (and holding)

At the end of 28 cycles, 50 μl of the reaction mix was removed; and the remaining reaction mix was run for an additional 10 cycles as outlined below:

Step 1 94° C. for 15 sec
Step 2 65° C. for 1 min

Step 3 68° C. for (10 min+15 sec)/cycle

Step 4 Repeat step 1–3 for 9 additional cycles

Step 5 72° C. for 10 min

A 5–10 μl aliquot of the reaction mixture is analyzed by electrophoresis on a low concentration (about 0.6–0.8%) agarose mini-gel to determine which reactions were successful in extending the sequence. Although all extensions potentially contain a full length gene, some of the largest products or bands are selected and cut out of the gel. Further purification involves using a commercial gel extraction method such as QIAQuick™ (QIAGEN Inc). After recovery of the DNA, Klenow enzyme is used to trim single-stranded, nucleotide overhangs creating blunt ends which facilitate religation and cloning.

After ethanol precipitation, the products are redissolved in 13 μl of ligation buffer. Then, 1 μl T4-DNA ligase (15 units) and 1 μl T4 polynucleotide kinase are added, and the mixture is incubated at room temperature for 2–3 hours or overnight at 16° C. Competent *E. coli* cells (in 40 μl of appropriate media) are transformed with 3 μl of ligation mixture and cultured in 80 μl of SOC medium (Sambrook J. et al, supra). After incubation for one hour at 37° C., the whole transformation mixture is plated on Luria Bertani (LB)-agar (Sambrook J. et al, supra) containing 2×Carb. The following day, 12 colonies are randomly picked from each plate and cultured in 150 μl of liquid LB/2×Carb medium placed in an individual well of an appropriate, commercially-available, sterile 96-well microtiter plate. The following day, 5 μl of each overnight culture is transferred into a non-sterile 96-well plate and after dilution 1:10 with water, 5 μl of each sample is transferred into a PCR array.

For PCR amplification, 15 μl of concentrated PCR reaction mix (1.33x) containing 0.75 units of Taq polymerase, a vector primer and one or both of the gene specific primers used for the extension reaction are added to each well. Amplification is performed using the following conditions:

Step 1 94° C. for 60 sec

Step 2 94° C. for 20 sec

Step 3 55° C. for 30 sec

Step 4 72° C. for 90 sec

Step 5 Repeat steps 2–4 for an additional 29 cycles

Step 6 72° C. for 180 sec

Step 7 4° C. (and holding)

Aliquots of the PCR reactions are run on agarose gels together with molecular weight markers. The sizes of the PCR products are compared to the original partial cDNAs, and appropriate clones are selected, ligated into plasmid and sequenced.

V Diagnostic Assays Using Kinase Specific Oligomers

In those cases where a specific disorder or disease (see definitions supra) is suspected to involve altered quantities of a particular kinase, oligomers may be designed to establish the presence and/or quantity of mRNA expressed in a biological sample. There are several methods currently being used to quantitate the expression of a particular molecule. Most of these methods use radiolabelled (Melby P. C. et al 1993 J Immunol Methods 159:235–44) or biotinylated (Duplaa C. et al 1993 Anal Biochem 229–36) nucleotides, coamplification of a control nucleic acid, and standard curves onto which the experimental results are interpolated. For example, phosphorylase B kinase deficiency may manifest as hepatomegaly which is inherited as either an X-linked or autosomal recessive trait or myoglobinuria whose inheritance is unknown.

Oligomers for phosphorylase B kinase are first used in quantitative PCR to establish a normal range for expression of phosphorylase B kinase. Then, these same oligomers are used with extracts of cells from patients with inherited phosphorylase B kinase deficiency. The information from such studies is used to define different inheritance patterns and to diagnose future patients displaying phosphorylase B kinase deficiency-like symptoms. In like manner, this same assay can be used to monitor progress of the patient as his/her physiological situation moves toward the normal range during therapy for the condition.

VI Kinases Kit

The kinases of the subject invention are used to produce a kinases kit for diagnosing disorders or diseases associated with altered kinase expression. This involves the designing a plurality of oligomers, one set of which is specific for each kinase or kinase regulatory sequence. Specificity in this case refers to sequence similarity, to the length of the nucleic acid molecule amplified, to cell or tissue type being screened or to the disorder or disease. These oligomers are combined with a biological sample obtained from a patient in a solution sufficient for PCR and amplified. The PCR products are examined first, to detect the expression of each kinase, and second to quantify the expression of each kinase. Kinase expression is compared with standard ranges for normal and abnormal expression. In the case(s) where kinase expression is altered, use of the kit has provided the physician with a named disorder or disease which can be treated or further investigated.

A further use of the oligomers from the kinases kit is in a diagnostic assay of example V (above) used to monitor patient response to drug therapy. Once the disease has been named and a therapy chosen, the oligomers specific to the patient's disease may be used periodically to monitor the efficacy of the chosen therapy. In this case, the specific oligomers are combined with a biological sample from the patient in a solution sufficient for PCR and amplified. The PCR product is quantified and compared with a normal standard and with the pretreatment profile of the patient. If the kinase expression is tending toward normal, the therapy may be considered effective; if the expression is even more abnormal, therapy should be discontinued and an alternative treatment instituted.

VII Sense or Antisense Molecules

Knowledge of the correct cDNA sequence of any particular kinase, its regulatory elements or parts thereof will enable its use as a tool in sense (Youssoufian H. and H. F. Lodish 1993) Mol Cell Biol 13:98–104) or antisense (Eguchi et al (1991) Annu Rev Biochem 60:631–652) technologies for the investigation of gene function. Oligonucleotides, from genomic or cDNAs, comprising either the sense or the antisense strand of the cDNA sequence can be used in vitro or in vivo to inhibit expression. Such technology is now well known in the art, and oligonucleotides or other fragments can be designed from various locations along the sequences.

The gene of interest can be turned off in the short term by transfecting a cell or tissue with expression vectors which will flood the cell with sense or antisense sequences until all copies of the vector are disabled by endogenous nucleases. Stable transfection of appropriate germ line cells or preferably a zygote with a vector containing the fragment will produce a transgenic organism (U.S. Pat. No. 4,736,866, 12 Apr. 1988), which produces enough copies of the sense or antisense sequence to significantly compromise or entirely eliminate normal activity of the particular kinase gene. Frequently, the function of the gene can be ascertained by observing behaviors such as lethality, loss of a physiological pathway, changes in morphology, etc. at the intracellular, cellular, tissue or organismal level.

In addition to using fragments constructed to interrupt transcription of the open reading frame, modifications of gene expression can be obtained by designing antisense sequences to promoters, enhancers, introns, or even to trans-acting regulatory genes. Similarly, inhibition can be achieved using Hogeboom base-pairing methodology, also known as "triple helix" base pairing.

VIII Expression of Kinases

Expression of the kinases may be accomplished by subcloning the cDNAs into appropriate vectors and transfecting the vectors into host cells. In some cases, the cloning vector previously used for the generation of the tissue library also provides for direct expression of kinase sequences in *E. coli.* Upstream of the cloning site, this vector contains a promoter for β-galactosidase, followed by sequence containing the amino-terminal Met and the subsequent 7 residues of β-galactosidase. Immediately following these eight residues is a bacteriophage promoter useful for transcription and a linker containing a number of unique restriction sites.

Induction of an isolated, transfected bacterial strain with IPTG using standard methods will produce a fusion protein corresponding to the first seven residues of β-galactosidase, about 5 to 15 residues which correspond to linker, and the peptide encoded within the kinase cDNA. Since cDNA clone inserts are generated by an essentially random process, there is one chance in three that the included cDNA will lie in the correct frame for proper translation. If the cDNA is not in the proper reading frame, it can be obtained by deletion or insertion of the appropriate number of bases by well known methods including in vitro mutagenesis, digestion with exonuclease III or mung bean nuclease, or oligonucleotide linker inclusion.

The kinase cDNA can be shuttled into other vectors known to be useful for expression of protein in specific hosts. Oligonucleotide linkers containing cloning sites as well as a stretch of DNA sufficient to hybridize to the end of the target cDNA (25 bases) can be synthesized chemically by standard methods. These primers can then used to amplify the desired gene fragments by PCR. The resulting fragments can be digested with appropriate restriction enzymes under standard conditions and isolated by gel electrophoresis. Alternatively, similar gene fragments can be produced by digestion of the cDNA with appropriate restriction enzymes and filling in the missing gene sequence with chemically synthesized oligonucleotides. Partial nucleotide sequence from more than one gene can be ligated together and cloned in appropriate vectors to optimize expression.

Suitable expression hosts for such chimeric molecules include but are not limited to mammalian cells such as Chinese Hamster Ovary (CHO) and human 293 cells, insect cells such as Sf9 cells, yeast cells such as *Saccharomyces cerevisiae*, and bacteria such as *E. coli.* For each of these cell systems, a useful expression vector may also include an origin of replication to allow propagation in bacteria and a selectable marker such as the β-lactamase antibiotic resistance gene to allow selection in bacteria. In addition, the vectors may include a second selectable marker such as the neomycin phosphotransferase gene to allow selection in transfected eukaryotic host cells. Vectors for use in eukaryotic expression hosts may require RNA processing elements such as 3' polyadenylation sequences if such are not part of the cDNA of interest.

Additionally, some of the kinase vectors may contain native promoters which will allow induction of gene expression in human cells such as the 293 line mentioned above. Other available promoters are host specific and may be specifically combined with the coding region of the kinase of interest. They include MMTV, SV40, and metallothionine promoters for CHO cells; trp, lac, tac and T7 promoters for bacterial hosts; and alpha factor, alcohol oxidase and PGH promoters for yeast. In addition, transcription enhancers, such as the rous sarcoma virus (RSV) enhancer, may be used in mammalian host cells. Once homogeneous cultures of recombinant cells are obtained through standard culture methods, large quantities of recombinantly produced peptide can be recovered from the conditioned medium and analyzed using methods known in the art.

IX Isolation of Recombinant KIN

KIN may be expressed as a recombinant protein with one or more additional polypeptide domains added to facilitate protein purification. Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals, protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp, Seattle, Wash.). The inclusion of a cleavable linker sequence such as Factor XA or enterokinase (Invitrogen) between the purification domain and the kin sequence may be useful to facilitate expression of KIN.

X Testing for Kinase Activity

The sequences in this application represent many different domains of different kinase families. These domains (and subdomains as detailed in the background of the invention) may be utilized: 1) individually for the production of antibodies, 2) in functional groups (eg. to span a membrane), and 3) as interchangable, usable parts of a chimeric kinase. The various partial cDNA sequences of this application represent the different kinase domains of the various families (Hardie G. and Hanks S., supra), and they may be recombined in numerous ways to produce chimeric nucleic acid molecules. For example, a known, full length kinase such as the human map kinase of this application (Seq ID No 45) may be used to swap related portions of the nucleic acid sequence, analogous to domains or subdomains of MAP kinase polypeptides. The chimeric nucleotides, so produced, may be introduced into prokaryotic host cells (as reviewed in Strosberg A. D. and Marullo S. (1992) Trends Pharma Sci 13:95–98) or eukaryotic host cells. These host cells are then employed in procedures to determine what molecules activate the kinase or what molecules are activated by a kinase. Such activating or activated molecules may be of extracellular, intracellular, biologic or chemical origin.

An example of a test system, in this case for protein tyrosine kinases, can be based on the interaction of protein tyrosine kinases with chemokine receptors (Taniguchi T. (1995) Science 268:251–255). These receptors are capable of activating a variety of nonreceptor protein tyrosine kinases when stimulated by an extracellular chemokine. C-X-C chemokines such as platelet factor 4, interleukin-8, connective tissue activating protein III, neutrophil activating peptide 2, are soluble activators of neutrophils.

A standard measure of neutrophil activation involves measuring the mobilization of Ca^{++} as part of the signal transduction pathway. The experiment involves several steps. First, blood cells obtained from venipuncture are fractionated by centrifugation on density gradients. Enriched populations of neutrophils are further fractionated on columns by negative selection using antibodies specific for other blood cells types. Next, neutrophils are transformed with an expression vector containing the kinase nucleic acid sequence of interest and preloaded fluorescent probe whose emission characteristics have been altered by Ca^{++} binding. Or in the alternative, the neutrophil is preloaded with the purified kinase of interest and fluorescent probe. Then, when the cells are exposed to an appropriate chemokine, the chemokine receptor activates the kinase which, in turn, initiates Ca^{++} flux. Ca^{++} mobilization is observed and measured using fluorometry as has been described in Grynkievicz G. et al (1985) J Biol Chem 260:3440, and McColl S. et al (1993) J Immunol 150:4550–4555, incorporated herein by reference.

XI Identification of or Production of Kinase Specific Antibodies

Purified KIN is used to screen a pre-existing antibody library or to raise antibodies. using either polyclonal or monoclonal methodology. For polyclonal antibody production, denatured peptide from the reverse phase HPLC separation is obtained in quantities up to 75 mg. This denatured protein can be used to immunize mice or rabbits using standard protocols; about 100 micrograms are adequate for immunization of a mouse, while up to 1 mg might be used to immunize a rabbit. In identifying mouse hybridomas, the denatured protein can be labelled and used to screen potential murine B-cell hybridomas for those which produce antibody. This procedure requires only small quantities of protein, such that 20 mg would be sufficient for labelling and screening of several thousand clones.

For monoclonal antibody production, the amino acid sequence, as deduced from translation of the cDNA, is analyzed to determine regions of high immunogenicity. Peptides comprising appropriate hydrophilic regions are expressed from recombinant cDNA or synthesized and used in suitable immunization protocols to raise antibodies. Selection of appropriate epitopes is described by Ausubel F. M. et al (supra). The optimal amino acid sequences for immunization are usually located at the C-terminus or N-terminus and in intervening, hydrophilic regions of the polypeptide which are likely to be exposed to the external environment when the protein is in its natural conformation.

Typically, selected oligopeptides, about 15 residues in length, are synthesized using an Applied Biosystems Peptide Synthesizer Model 431A using fmoc-chemistry and coupled to keyhole limpet hemocyanin (KLH, Sigma) by reaction with M-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS; Ausubel F. M. et al, supra). If necessary, a cysteine may be introduced at the N-terminus of the peptide to permit coupling to KLH. Rabbits are immunized with the peptide-KLH complex in complete Freund's adjuvant. The resulting antisera are tested for antipeptide activity by binding the peptide to plastic, blocking with 1% bovine serum albumin, reacting with antisera, washing and reacting with labelled, affinity purified, specific goat anti-rabbit IgG.

Hybridomas may also be prepared and screened using standard techniques. Hybridomas of interest are detected by screening with labelled KIN to identify those fusions producing the monoclonal antibody with the desired specificity. In a typical protocol, wells of plates (FAST; Becton-Dickinson, Palo Alto, Calif.) are coated during incubation with affinity purified, specific rabbit anti-mouse (or suitable anti-species Ig) antibodies at 10 mg/ml. The coated wells are blocked with 1% BSA, washed and incubated with supernatants from hybridomas. After washing the wells are incubated with labelled KIN at 1 mg/ml. Supernatants with specific antibodies bind more labelled KIN than is detectable in the background. Then clones producing specific antibodies are expanded and subjected to two cycles of cloning at limiting dilution. Cloned hybridomas are injected into pristane-treated mice to produce ascites, and monoclonal antibody is purified from mouse ascitic fluid by affinity chromatography on Protein A. Monoclonal antibodies with affinities of at least 10^8/M, preferably 10^9 to 10^{10} or stronger, will typically be made by standard procedures as described in Harlow and Lane (1988) Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.; and in Goding (1986) Monoclonal Antibodies: Principles and Practice, Academic Press, New York City, both incorporated herein by reference.

XII Diagnostic Assays Using KIN Specific Antibodies

Particular KIN antibodies are useful for investigation of various disorders or diseases which may be characterized by differences in the amount or distribution of KIN. Given the usual role of the kinases, KIN might be expected to be upregulated (or downregulated) in its involvement in activation of signal cascades.

Diagnostic assays for KIN include methods utilizing the antibody and a reporter molecule to detect KIN in human body fluids, membranes, cells, tissues or extracts thereof. The antibodies of the present invention may be used with or without modification. Frequently, the antibodies will be labelled by joining them, either covalently or noncovalently, with a substance which provides for a detectable signal. A wide variety of reporter molecules and conjugation techniques are known and have been reported extensively in both the scientific and patent literature. Suitable reporter molecules or labels include those radionuclides, enzymes, fluorescent, chemi-luminescent, or chromogenic agents previously mentioned as well as substrates, cofactors, inhibitors, magnetic particles and the like. Patents teaching the use of such labels include U.S. Pat. Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241. Also, recombinant immuno-globulins may be produced as shown in U.S. Pat. No. 4,816,567, incorporated herein by reference.

A variety of protocols for measuring soluble or membrane-bound KIN, using either polyclonal or monoclonal antibodies specific for the protein, are known in the art. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA) and fluorescent activated cell sorting (FACS). A two-site monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering epitopes on KIN is preferred, but a competitive binding assay may be employed. These assays are described, among other places, in Maddox, D. E. et al (1983, J Exp Med 158:1211).

XIII Purification of Native KIN Using Antibodies

Native or recombinant protein kinases can be purified by immunoaffinity chromatography using antibodies specific for that particular KIN. In general, an immunoaffinity column is constructed by covalently coupling the anti-KIN antibody to an activated chromatographic resin.

Polyclonal immunoglobulins are prepared from immune sera either by precipitation with ammonium sulfate or by purification on immobilized Protein A (Pharmacia Biotech). Likewise, monoclonal antibodies are prepared from mouse ascites fluid by ammonium sulfate precipitation or chromatography on immobilized Protein A. Partially purified immunoglobulin is covalently attached to a chromatographic resin such as CnBr-activated Sepharose (Pharmacia Biotech). The antibody is coupled to the resin, the resin is blocked, and the derivative resin is washed according to the manufacturer's instructions.

Such immunoaffinity columns may be utilized in the purification of KIN by preparing a fraction from cells containing KIN in a soluble form. This preparation may be derived by solubilization of whole cells or of a subcellular fraction obtained via differential centrifugation (with or without addition of detergent) or by other methods well known in the art. Alternatively, soluble KIN containing a signal sequence may be secreted in useful quantity into the medium in which the cells are grown.

A soluble KIN-containing preparation is passed over the immunoaffinity column, and the column is washed under conditions that allow the preferential absorbance of KIN (eg, high ionic strength buffers in the presence of detergent). Then, the column is eluted under conditions that disrupt antibody/KIN binding (eg, a buffer of pH 2–3 or a high concentration of a chaotrope such as urea or thiocyanate ion), and KIN is collected.

XIV Drug Screening

This invention is particularly useful for screening therapeutic compounds by using binding fragments of KIN in any of a variety of drug screening techniques. The molecules to be screened may be of extracellular, intracellular, biologic or chemical origin. The peptide fragment employed in such a test may either be free in solution, affixed to a solid support, borne on a cell surface or located intracellularly. One may measure, for example, the formation of complexes between KIN and the agent being tested. Alternatively, one can examine the diminution in complex formation between KIN and a receptor caused by the agent being tested.

Methods of screening for drugs or any other agents which can affect signal transduction comprise contacting such an agent with KIN fragment and assaying for the presence of a complex between the agent and the KIN fragment. In such assays, the KIN fragment is typically labelled. After suitable incubation, free KIN fragment is separated from that present in bound form, and the amount of free or uncomplexed label is a measure of the ability of the particular agent to bind to KIN.

Another technique for drug screening provides high throughput screening for compounds having suitable binding affinity to the KIN polypeptides and is described in detail in European Patent Application 84/03564, published on Sep. 13, 1984, incorporated herein by reference. Briefly stated, large numbers of different small peptide test compounds are synthesized on a solid substrate, such as plastic pins or some other surface. The peptide test compounds are reacted with KIN fragment and washed. Bound KIN fragment is then detected by methods well known in the art. Purified KIN can also be coated directly onto plates for use in the aforementioned drug screening techniques. In addition, non-neutralizing antibodies can be used to capture the peptide and immobilize it on the solid support.

This invention also contemplates the use of competitive drug screening assays in which neutralizing antibodies capable of binding KIN specifically compete with a test compound for binding to KIN fragments. In this manner, the antibodies can be used to detect the presence of any peptide which shares one or more antigenic determinants with KIN.

XV Identification of Molecules Which Interact with KIN

The inventive purified KIN is a research tool for identification, characterization and purification of interacting, signal transduction pathway proteins. Appropriate labels are incorporated into KIN by various methods known in the art and KIN is used to capture soluble or interact with membrane-bound molecules. A preferred method involves labeling the primary amino groups in KIN with ^{125}I Bolton-Hunter reagent (Bolton, A. E. and Hunter, W. M. (1973) Biochem J 133:529). This reagent has been used to label various molecules without concomitant loss of biological activity (Hebert C. A. et al (1991) J Biol Chem 266:18989–94; McColl S. et al (1993) J Immunol 150:4550–4555). Membrane-bound molecules are incubated with the labelled KIN molecules, washed to removed unbound molecules, and the KIN complex is quantified. Data obtained using different concentrations of KIN are used to calculate values for the number, affinity, and association of KIN with the signal transduction complex.

Labelled KIN fragments are also useful as a reagent for the purification of molecules with which KIN interacts, specifically including inhibitors. In one embodiment of affinity purification, KIN is covalently coupled to a chromatography column. Cells and their membranes are extracted, KIN is removed and various KIN-free subcomponents are passed over the column. Molecules bind to the column by virtue of their KIN affinity. The KIN-complex is recovered from the column, dissociated and the recovered molecule is subjected to N-terminal protein sequencing. This amino acid sequence is then used to identify the captured molecule or to design degenerate oligomers for cloning its gene from an appropriate cDNA library.

In an alternate method, monoclonal antibodies raised against KIN fragments are screened to identify those which inhibit the binding of labelled KIN. These monoclonal antibodies are then used in affinity purification or expression cloning of associated molecules. Other soluble binding molecules are identified in a similar manner. Labelled KIN is incubated with extracts or other appropriate materials derived from rheumatoid synovium. After incubation, KIN complexes (which are larger than the lone KIN fragment) are identified by a sizing technique such as size exclusion chromatography or density gradient centrifugation and are purified by methods known in the art. The soluble binding protein(s) are subjected to N-terminal sequencing to obtain information sufficient for database identification, if the soluble protein is known, or for cloning, if the soluble protein is unknown.

XVI Use and Administration of Antibodies or Other Inhibitory Molecules

Antibodies, inhibitors, receptors or antagonists of KIN fragments (or other treatments to limit signal transduction, TST), can provide different effects when administered therapeutically. TSTs will be formulated in a nontoxic, inert, pharmaceutically acceptable aqueous carrier medium preferably at a pH of about 5 to 8, more preferably 6 to 8, although the pH may vary according to the characteristics of the antibody, inhibitor, or antagonist being formulated and the condition to be treated. Characteristics of TSTs include solubility of the molecule, half-life and antigenicity/immunogenicity; these and other characteristics may aid in defining an effective carrier. Native human proteins are preferred as TSTs, but organic or synthetic molecules resulting from drug screens may be equally effective in particular situations.

TSTs may be delivered by known routes of administration including but not limited to topical creams and gels; transmucosal spray and aerosol; transdermal patch and bandage; injectable, intravenous and lavage formulations; and orally administered liquids and pills particularly formulated to resist stomach acid and enzymes. The particular formulation, exact dosage, and route of administration will be determined by the attending physician and will vary according to each specific situation.

Such determinations are made by considering multiple variables such as the condition to be treated, the TST to be administered, and the pharmacokinetic profile of the particular TST. Additional factors which may be taken into account include disease state (e.g. severity) of the patient, age, weight, gender, diet, time and frequency of administration, drug combination, reaction sensitivities, and tolerance/response to therapy. Long acting TST formulations might be administered every 3 to 4 days, every week, or once every two weeks depending on half-life and clearance rate of the particular TST.

Normal dosage amounts may vary from 0.1 to 100,000 micrograms, up to a total dose of about 1 g, depending upon the route of administration. Guidance as to particular dosages and methods of delivery is provided in the literature. See U.S. Pat. No. 4,657,760; 5,206,344; or 5,225,212. Those skilled in the art will employ different formulations for different TSTs. Administration to cells such as nerve cells necessitates delivery in a manner different from that to other cells such as vascular endothelial cells.

It is contemplated that disorders or diseases which trigger defensive signal transduction may precipitate damage that is treatable with TSTs. These disorders or diseases may be specifically diagnosed by the tests discussed above, and such testing should be performed in cases where physiologic or pathologic problems are suspected to be associated with abnormal signal transduction.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention which are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

TABLE 1

Clone	Library	GenBank/SwissProt Identifier, Name
297	U937	P00540 Mouse protooncogene ser/thr kinase
1622	U937	HUMCLK3B clk3 gene product
10007	THP-1 Phorbol LPS	HSPLK1 protein kinase
12702	THP-1 Phorbol LPS	RATSGPK ser/thr kinase
23789	Inflamed Adenoid	CHKFRNK chicken tyr kinase
35652	HUVEC	KEK5 Chicken Y kinase receptor
35855	HUVEC	HUMANBTK37 tyr kinase
40194	T + B Lymphoblast	KRB1 VARV Variola virus protein kinase
42170	T + B Lymphoblast	HSU09564 serine kinase
46081	Corneal Stroma	YSCKIN1 yeast protein kinase
46651	Corneal Stroma	CDK4, P11802
53840	Fibroblast	HSDAPK, Death-associated protein kinase
54065	Fibroblast	SCPROKIN 1 yeast 35.6 kD
56494	Fibroblast	KLMC RAT, myosin light chain kinase
58029	Skeletal Muscle	ATHCTRIA 1 *A. Thaliana* Y kinase receptor
64663	Placenta	KIN3 Yeast protein kinase P22209
67967	HUVEC Sheer Stress	YAK1 Yeast protein kinase
68963	HUVEC Sheer Stress	KATK Human Y kinase
71904	Placenta	KIN3 P22209SwP
75289	THP-1 Phorbol	H5U08023 Avian retrovirus rp130
81865	Rheumatoid Synovium	SNF1 Yeast C catabolite derepressing
82056	HUVEC Sheer Stress	P34314 *C. elegans* ser/thr kinase
108485	AML Blast	KAPA Pig cAMP-dependent protein kinase
114973	Testis	CC2B ARATH Mouse-ear cress cdc
118591	Skeletal Muscle	PBO192 mixed lineage kinase 1
119819	Skeletal Muscle	H5U09564 ser kinase
120376	Skeletal Muscle	U01064 Y kinase
132750	Bone Marrow	MLK2 mixed lineage kinase 2
140052	T Lymphocyte	G-protein coupled receptor kinase
146392	T Lymphocyte	SCYAK1 Yeast Yak1 kinase
156108	THP-1 Phorbol LPS	U01064 Dictyostelium Y kinase
173627	Bone Marrow	MMU14166 Kiz
181971	Placenta	HUMTKR Y kinase receptor
182538	Placenta	HSNEK2R kinase
184416*	Cardiac Muscle	KPKS Human proto-oncogene Ser/Thr kinase
191283	Rheumatoid Synovium	RATSGPK Ser/Thr kinase
192268	Rheumatoid Synovium	ATHAPK1A Ser/Thr kinase
214915	Stomach	XLMPK2K Map kinase
223163	Pancreas	TGF-β receptor ser/thr kinase
237002	Small Intestine	P16227 Mouse Y kinase blk
239990	Hippocampus	SHC Human transforming protein
240142	Hippocampus	HSNEK2R
275781	Testes	BOVCKIA casein kinase
285465	Eosinophils	DDIMLCK myosin light chain kinase

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 45

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 526 base pairs

-continued

(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: U937
(B) CLONE: 297

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
ACAAGGGTTG  TAATTAAAGG  CGATTTTGAA  ACAATTAAAA  TCTGTGATGT  AGGAGTCTCT    60
CTACCACTGG  ATGAAAATAT  GACTGTGACT  GACCCTGAGG  CTTGTTACAT  TGGCACAGAG   120
CCATGGAAAC  CCAAAGAAGC  TGTGGAGGAG  AATGGTGTTA  TTACTGCAAG  GCAGACATAT   180
TTGCCTTTGG  CTTACTTTGT  GGGAAATGAT  GACTTTATCG  ATTCCACACA  TTAATCTTTC   240
AAATGATGAT  GATGATGAAG  TAAAAACTTT  TTGATGAAAA  GTAATTTTGA  TGTTGAAGCA   300
TTACTATGCA  AGCCCTTTGG  ACCTAAGGCC  ACCCTATTTT  AATATTGGAG  GACCTTGGTG   360
AATCATACCC  AGGAAGGTAA  TTTGACCTCT  TCTCTGATCA  CCCTTATTGA  AGCCCCCAAG   420
CACCCTTCTT  GTGACAATTT  TAGGTTGGAC  CAGTTGCTTT  GGGCCAACTT  AACTAAAGTT   480
GTTCGAAAAA  CTTTTTTCCA  AAAATTTCCA  TAGGCCTCCC  AAGTTT                   526
```

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 378 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: U937
(B) CLONE: 1622

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
AGAACACCAC  ATCCGAGTGG  CTGACTTTGG  CAGTGCCACA  TTTGACCATG  AGCACCACAC    60
CACCATTGTG  GCCACCCGTC  ACTATCGCCG  CCTGAGGTGA  TCCTTGAGCT  GGGCTGGGCA   120
CAGCCTGGTG  ACGTCTGGGC  ATTGGCTGCA  TTCTCTTTGA  GTACTACCGG  GGCTTCACAC   180
TCTTCCAGAC  CCACGAAAAC  CGAGAGCACC  TGGTGATGAT  GGAGAAGATC  CTAGGGCCCA   240
TCCCATCACA  CATGATCCAC  CGTACCAGGA  AGCAGAATAT  TTCTACAAAG  GGGGCCTAGT   300
TTGGGATGGA  CAGCTCTTAC  GGCCGGTATG  TAAGGGACTC  AAACCTTTAA  GGTTCATGTT   360
CAAGCTTCCT  GGGAAGTG                                                     378
```

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 326 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: THP-1 Phorbol LPS
(B) CLONE: 10007

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

```
GGGCTGGCAG  CCCGGTTGGA  GCCTCCGGAG  CAGAGGAAGA  AGACCATCTT  GGCACCCCCA    60
```

-continued

```
ACTATGTGGC TCCAGAAGTG CTGCTGAGAC AGGGCCACGG CCCTGAGGCG GATGTATGGT   120
CACTGGGCTG TGTCATGTAC ACGCTGCTCT GCGGGACCCT CCCTTTGAGA CGGCTGACCT   180
GAAGGAGACG TACCGCTGCA TCAAGAAGGT TCACTACAAC GGTGCCTGCC AGCTCTTAAT   240
GCCTGCCCGA GTCCTTGGCC GCAATCCTTC GGGCCTTAAC CCGAGAACCG GCCCTCTATT   300
GACAGATCCT TGCGGCAATT AACTTT                                        326
```

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 257 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: THP-1 Phorbol LPS
(B) CLONE: 12702

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
CCGCAAGACA CCTCCTGGAG GGCCTCCTGA GAAGGACAGG CAAAGGGCTG GGCCAAGGAT    60
GACTTCATGG AGATTAAGAG TCATGTTTCT TCTCCTTAAT TAACTGGGAT GATCTCATTA   120
ATAAGAAGAT TACTCCCCCT TTTACCCAAA TGTGAGTGGG CCCAACGCCT ACGGACTTTG   180
CCCCGAGTTT ACGAAGAGCC TTCCCCAATC CATTGGAAGT CCCCTGAAAG GTCCTATACA   240
AGTCAGTTAA GGAAGTT                                                  257
```

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 252 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Inflamed Adenoid
(B) CLONE: 23789

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

```
GTGAAGAATG TGGGGCTGAC CCTCGGAAGT CATCGGGAGC GTGGATGATC TCCTGCCTTC    60
CTTGCCGTCA TCTCACGGAC AGAGATCGAG GGCACCCAGA AACTGCTCAA CAAAGACCTG   120
GCAGAGCTCA TCAACAAGAT GCGCTGGCGC AAGAACGCGT GACCTCCCTG TAGGAGTAAG   180
AGGCAGATCT GACGGTTCAC AACCCTGGCT GTGACGCAAG AACCTCTTAC GTGTGCCAGG   240
CCCAAAGTTC TG                                                       252
```

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 255 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Huvec
(B) CLONE: 35652

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

-continued

```
CAAAATCGTG GCCCGGAGAA TGGCGGGGCC TCAACCCTCT CCTGGACCAG CGGCAGCTCA     60
CTACTCAGCT TTTGGCCTGT GGGCGAGTGG CTTCGGGCCA TCAAAATGGG AAGATACGAA    120
GAAAGTTTCG CAGCCGCTGG CTTTGGCTCC TTCAGCTGGT CAGCCAGATC TCTGCTGAGG    180
ACCTGCTCCG AATCGAGTCA CTCTGGCGGG ACACCAGAAG AAAATTTGGC CAGTTCCAGC    240
ACATGAGTCC CAGGT                                                     255
```

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 238 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Huvec
(B) CLONE: 35855

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
GAATACCCCA TATACATAGT GACTGATATA TAAGCAATGG CTGCTTGCTG AATACCTGAG     60
GAGTCACGGA AAAGGCTTAA CCTTCCCAGT CTTAGAAATG TGCTACGATG TCTGTAAGGC    120
ATGGCCTTCT TGGAGAGTCA CCAATTCATA CACCGGGCTT GGCTGCTCGT AACTGCTTGG    180
TGGACAGAGA TCTCTGTGTG AAAGTTCTCC ATTTGGATGA CAAGGTATGT TCTTGATG      238
```

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 261 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: T+B Lymphoblast
(B) CLONE: 40194

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
AAACAACTTG ATTATTTAGG AATTCCTCTG TTTTATGGAT CTGGTCTGAC TGAATTCAAG     60
GGAAGAAGTT ACAGATTTAT GGTAATGGAA AGACTAGGAA TAGATTTACA GAAGATCTCA    120
GGCCAGAATG GTACCTTTAA AAAGTCAACT GTCCTGCAAT TAGGATCCGA ATGTTGGATG    180
TACTGGAATA TATACATGAA AATGAATATG TTCATGGTGA TATAAAAGCA GCAAATCTAC    240
TTTTGGGTTA CAAAAATCCT T                                              261
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 242 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: T+B Lymphoblast
(B) CLONE: 42170

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:9:

-continued

```
TAAGAAACCT GAAGATCGAG CCACTGCTGA AGAATGTCTA AAGCACCCCT GGTTGACACA    60

GAGCAGTATT CAAGAGCCTT CTTTCAGGAT GGAAAAGGCA CTAGAAGAAG CAAATGCCCT   120

CCAAGAAGGT CATTCTGTGC CTGAAATTAA TTCGGATACC GACAAATCAG AAACCGAGGA   180

ATCCATTGTA ACCGAAGAGT TAATTGTAGT TACTTCATAT ACTCTAGGGC AATGCAGACA   240

GT                                                                  242
```

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 222 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Corneal Stroma
(B) CLONE: 46081

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:10:

```
GCAAAGGACA GTCCGCCGAG GTGCTCGGTG GAGTCATGGC ATTCCCTTTT GGAAGACTGG    60

CCTTGGTGCA AACCCTGGAG AAGGTGCCTA TGGAGAAGTT CAACTTGCTG TAAATAGAGT   120

AACTAAGAAG CAGTCGCAGT GAAGATTTAG ATATAAGCGT GCCGTAGACT GTCCCGAAAA   180

TATTAAGTAG ATCTGTATCA ATAAAATGCT AATCATGAAA TT                      222
```

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 225 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Corneal Stroma
(B) CLONE: 46651

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:11:

```
ATGCTCCGCC AGTGAGAAGG GCGGCTGCCT GAGCGCCTCA CCAGTCCTCA TCACCCAGAT    60

CCTGTGGCTT TGAGACACCT TCACTTAAGA ACATTTGCCA CTTGACTTAA ACCAGAAACG   120

TGTTTTGTGG CATCAGCAGA CCCTTTCTCA GGTAAGTTGT GCTTTGCTTT TAGCATACGT   180

GAGAAGTTGT TCCGCTCCAT TTTGTGGGAC GTCTTTCTTT CCTTG                   225
```

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 256 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Fibroblast
(B) CLONE: 53840

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:12:

```
CAGCGCCTTA CATCTCGCAG CCAAGAACAG CCACCATGAA TGCATCAGGA AGCTGCTTCA    60

TCTAAATGCC CAGCCGAAAG TTTTGACAGC TCTGGGAAAA CAGCTTTACA TTATGCAGCG   120
```

-continued

```
GCTCAGGGCT GCCTTCAAGC TGTGCAGATT CTTGCGAACA CAAGAGCCCC ATAAACCTCA   180

AAGATTTGGA TGGGAATATA CCGCTGCTGC TTGCTGTACA AAATGGTCAC AGTGAGATCT   240

GTCACTTTTC CTGGTC                                                   256
```

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 240 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Fibroblast
(B) CLONE: 54065

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:13:

```
GTTGACATCT GGTCCCTGGG CATATGGCCA TCGAAATGAT TGAAGGGGAG CCTCATACCT    60

CAATGAAAAC CCTTGAGAGC CTTGTACCTC ATTGCCACCA ATGGGACCCC AGAACTTCAG   120

AACCCAGAGA AGCTGTCAGC TATCTTCCGG GACTTTCTGA ACCGCTGTCT CGAGATGGAT   180

GTGGAGAAGA GAGGTTCAGC TAAAGAGCTG CTACAGCATC AATTCCTGAA GATTGCCAAT   240
```

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 195 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Fibroblast
(B) CLONE: 56494

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:14:

```
AACAGTGAAG AGCTCCGAGA AATTATGGGT ACCCTGATAT GTGGCTCCTG AAATTTAGTT    60

ATGATCCTAT AAGCATGGCA ACAGATATTG GAGCATTGGA GTGTTAACAT ATGTCATGCT   120

TACAGGAATA TCACCTTTTT AGGCAATGAT AAACAAGAAA CATTCTTAAA CATCTCACAG   180

ATGATTTTAA GTTAT                                                    195
```

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 207 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Skeletal Muscle
(B) CLONE: 58029

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:15:

```
GGAGTGTTTA TCGAGCCAAA TGGATATCAC AGGACAAGGA GGTGGCTGTA AAGAAGCTCC    60

TCAAAATAGA GAAAGAGGCA GAAATACTCA GTGTCCTCAG TCACAGAAAC ATCATCCAGT   120

TTTATGGAGT AATTTTGAAC CTCCCAACTA TGGCATTGTC ACAGAATATG CTTCTTGGGT   180
```

-continued

```
CACTCTATGA TTACATTAAC AGTACAA                                           207
```

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 184 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Placenta
(B) CLONE: 64663

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:16:

```
CGGGGTGGTA AAACTTGGAG ATCTTGGGAT TGGCGGTTTT AGCTCAAAAA CCACAGCTGC    60

ACATTCTTTA GTTGGTACGC CTATTCATGT TCCAGAGGAT ACAGAAATGG ATACAACTTC   120

AAATCTCATC TGGTCTCTTG GCTGTCTACT ATATGGATGG CTGCATTACA AAGTCCTTTC   180

TATG                                                                 184
```

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 206 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: HUVEC Sheer Stress
(B) CLONE: 67967

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:17:

```
TGAATTGCTG AGCATAGACC TTTATGAGCT GATTAAAAAA AATAAGTTTC AGGTTTTAGC    60

GTCCAGTTGG TACGCAAGTT TGCCCAGTCC ATCTTGCAAT CTTTGGTGCC CTCCACAAAA   120

TAAGATTATT CACTGCGATC TGAGCCAGAA AACATTCTCC TGAAACACCA CGGGCGCAGT   180

TCAACCAAGG TCATTGACTT TGGGTT                                         206
```

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 268 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: HUVEC Sheer Stress
(B) CLONE: 68963

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:18:

```
GGGAAGTGGC CAGTTTGGAG TGGTCAGCTG GGCAAGTGGA AGGGGCAGTA TGATGTTGCT    60

GTTAAGATGA TCAAGGAGGG CTCCATGTCA GAAGATGAAT TTTTCAGGAG GCCCAGACTA   120

TATGAAACTC AGCCATCCCA AGCTGGTTAA ATTCTATGGA GTGTGTTAAA GGATTACCCC   180

ATATACATGT GACTAATATA TAGCAATGCT TGCTTTTCTG AATTACCTGG GGAGTCACGG   240

AAAAAGGACT TTTAACCCTT CCCGCTTG                                       268
```

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 224 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Placenta
(B) CLONE: 71904

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:19:

CCTGGGGTGG TAAAACTTGG AGACTTGGCT TGGCCGGTTT TCCACCTCAA AAACCACAGC 60

TGCACATCCT TTAGTTGGTA CGCCTTATTA CATGTTCCAG AGAGATACAT GAAAATGGAT 120

ACAACTCAAA CTGACATCTG GCCTTTGGCT GTTACTATAT GAATGGCTGC TTACAAAGCC 180

TTCCTATGGT GACAAAATGA TTTTACTCAT TGTGTAAGAG ATAG 224

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 195 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: THP-1 Phorbol
(B) CLONE: 75289

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:20:

GCGGGGAATG ACTCCCTATC CTGGGGTCCA GAACCATGAG ATGTATGATA TCTTCTCCAT 60

GGCCACAGGT TGAAGCAGCC CGAAGACTGC CTGGTGAACT GTATGAAATA ATGTACTCTT 120

GCTGGAGAAC CGATCCCTTA GACCGCCCCA CCTTTTCATA TTGAGGCTGC AGCTAGAAAA 180

ACTCTTAGAA AGTTT 195

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 219 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Rheumatoid Synovium
(B) CLONE: 81865

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:21:

CACACGAGAA GCAGAAACAC GACGGGCGGG TAAGATCGGC CACTACATTC TGGTGACACG 60

CTGGGGGTCG GCACCTTCGG CAAAGTGAAG GTTGGCAAAC ATGATTGACT GGCATAAAGT 120

AGCTGTAAGA TACTCATCGA CAGAAGATTC GGAGCCTTGA TGTGGTAGGA AAAATCCCAG 180

GAAATTCAGA ACCTCAAGCT TTTCAGGCAT CCTCATATA 219

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 181 base pairs
(B) TYPE: nucleic acid

-continued

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: HUVEC Sheer Stress
(B) CLONE: 82056

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:22:

```
CCACCAAAGA TCTCAAATAA AGTTGATGTG TGGTCGGTGG GTGTATCTCT ATCAGTGTCT    60

TTATGGAAGG AAGCCTTTTG GCCATAACCA GTCTCAGCAA GACATCCTAC AAGAGAATAC   120

GATTTTAAAG CTACTGAAGT GCAGTTCCCG CCAAAGCCAG TAGTAACACC TGAAGCAAAG   180

G                                                                   181
```

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 218 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: AML Blast
(B) CLONE: 108485

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:23:

```
TATGGTTATA TGGAAGAGAA TGTGACTGGT GGTCGGTTGG GGTATTTTTA TACGAAATGC    60

TTGTAGGTGA TACACCTTTT TATGCAGATT CTTTGGTTGG AACTTACAGT AAAATTATGA   120

ACCATAAAAA TTCACTTACC TTTCCTGATG ATAATGACAT ATCAAAAGAA GCAAAAAACC   180

TTATTTGTGC CTTCCTTACT GACAGGGAAG TGAGGTTA                           218
```

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 264 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Testis
(B) CLONE: 114973

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:24:

```
GACGGTGGCC ATTTGACATG TGGAGCCTGG GTGCATCACG GTGGAGTTGT ACACGGGCTA    60

CCCCCTGTTC CCCGGGAGAA TGAGGTGGAG CAGCTGGCCT GCATCATGGA GGTGCTGGGT   120

CTGCCGCCAG CCGGCTTCAT TCAGACAGCC TCCAGGAGAC AGACATTCTT TGATTCCAAA   180

GGTTTTCCTA AAAATATAAC CACAACCAGG GGAAAAAAAG ATTCCAGATT CCAAGGGCCC   240

TCACGGATTG GTGCTGAAAA AACT                                          264
```

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 236 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Skeltal Muscle
(B) CLONE: 118591

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:25:

```
GACTGAGGAC ACTGAAACAT CATCCAGTTT TATGGAGTAA TTCTTGAACC TCCCAACTAT    60
GGCATTGTCA CAGAATATGC TTCTCTGGGA TCACTCTATG ATTACATTAA CAGTAACAGA   120
AGTGAGGAGA TGGATATGGT CACATTATGA CCTGGGCCAC TGATGTAGCC AAAGGAATGC   180
ATTATTTACA TATGGGGCTC CTGTCAAGGT GATTCACAGA GACCTCAAGT CAAGGA       236
```

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 200 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Skeltal Muscle
(B) CLONE: 119819

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:26:

```
CCTGCATGGC CTTCGAGCTG GCCACTGGTG ACTACCTGTT CGAGCCGCAT TCTGGAGAAG    60
ACTACAGTCG TGATGAGGGT AAGGGGTGAG GGCTCTGGGC TCAGCCTCCC GGCCTCCCGG   120
CCTGCCTGCC CCCAACCTCC TCTTTTGCCC ACAGACCACA TCGCTCACAT AGTGGAGCTT   180
CTGGGGGACA TCCCCCCAGC                                               200
```

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 217 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Skeletal Muscle
(B) CLONE: 120376

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:27:

```
GATTACAAGT AGCTTGGTTG TAGTGGAAAA AAACGAGAGA TTAACCATTC CAAGCAGTTG    60
CCCCAGAAGT TTTGCTGAAC TTTACATCAG TTTGGGAAGC TGATGCCAAG AAACGGCCAT   120
CATTCAAGCA AATCATTTCA ATCCTGGGTC CATGTCAAAT GACACGAGCC TTCCTGCAAG   180
TGTAACTCAT TCCTACACAA CAAGGCGGAG TGGAGGT                            217
```

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 156 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Bone Marrow
(B) CLONE: 132750

-continued

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:28:

GTAGATTTGA CTCTGTTGTT TTCTCTCGTA GTTCCCAAAC TCATGGAAGT CTGTTTTTAT 60

CAATATGATG TAAAGTCTGA AATATACAGC TTTGGAATCG TCCTCTGGGA AATCGCCACT 120

GGAGATATCC CGTTTCAAGG CTGTAATTCT GAGAAG 156

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 224 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: T Lymphocyte
(B) CLONE: 140052

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:29:

TGTAAATAAG GCCCTTCTCC ACTTGACTTC AGGCAGCAGA TTGTCTAGAA GCCTAAGGAC 60

AGCAATTTCT CTGACAAGAC AAAGTAGATA TTTTATACCA GGGGTTGGCA AACTACTGCC 120

CACGGGCCGA ATTTGGCCCA GTCTGTTTTT GTATGGTGCA AACTAAAAAT GATTTTTACA 180

TTTTTAAAGA GTTATAAAAG AAAAAAATAT GTGGTCTGTG AAAT 224

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 198 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: T Lymphocyte
(B) CLONE: 146392

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:30:

TTTTCTTTGT GTTTTTTTTT GTTCCAGTTT ATTTTAAATG CATATTTTAG TTGATTGCTT 60

TTTTAAAAAG CCCCCTCTGG CCTCCTGATT CCAGCTAGTG TCAGCAGTGG GATACCTGCG 120

CTTGAAGGAC ATCATCCACC GTGACATCAA GGATGAGAAC ATCGTGATCG CCGAGGACTT 180

CACAATCAAG CTGATAGT 198

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 210 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: THP-1 Phorbol LPS
(B) CLONE: 156108

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:31:

TGAAAACTAT GAACCTGGAC AAAAATCAAG GGCCAGTATC AAGCACGATA TATATAGCTA 60

TGCAGTTATC ACATGGGAAG TGTTATCCAG AAAACAGCCT TTTGAAGATG TCACCAATCC 120

-continued

```
TTTGCAGATA ATGTATAGTG TGTCACAAGG ACATCGACCT GTTATTAATG AAGAAAGTTT    180

GCCATATGAT ATACCTCACC GAGCACGTAT                                     210
```

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 202 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Bone Marrow
(B) CLONE: 173627

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:32:

```
AGAAGATCGG GGCCGGCTTC TTCTCTGAGG TCTACAAGGT TCGGCACCGA CAGTCAGGGC     60

AAGTATGGTG CTGAAGATGA ACAAGCTCCC CAGTAACCGG GGCAACACAC TACGGGAAGT    120

GCAGCTGATG AACCGGCTCA GGCACCCCAA CATCCTAAGG TTCATGGGAG TCTGTGTGCA    180

CCAGGGACAG CTGCACGCTC TT                                             202
```

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 222 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Placenta
(B) CLONE: 181971

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:33:

```
CGTTTTTGGA GGGTTCACAC CTGTCCCTTT CAAATGCTGG CGCTTTCACA CACTCCTTCT     60

CTCCTGCCAG CACCTTCTGG TCTCAGGAGC ATTGCAGGAT GTTGTGTGAG TAAGTATGGG    120

AGACACTTTA GTATGGCTTT TTTCAGCTTA GCCTCCTGTT ATCAGAGAGC AGTCTCTTTC    180

AGTGTCAAGG TTTGAGTACT AGATGGTGGA GAAAGCCTGT TT                       222
```

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 192 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Placenta
(B) CLONE: 182538

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:34:

```
CTTGGGGTGG TAAAACTTGG AGATCTTGGG CTTGGCCGGT TTTTCAGCTC AAAAACCACA     60

GCTGCACATT CTTTAGTTGG TACGCCTTAT TACATGTCTC CAGAGAGAAT ACATGAAAAT    120

GGATACAACT TCAAATCTGA CATCTGGTCT CTTGGCTGTC TACTATATGA GATGGCTGCA    180

TTACAAAGTC CT                                                        192
```

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 152 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Cardiac Muscle
(B) CLONE: 184416

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:35:

```
CTATGGAAGG CCGCTGGCAG GGCAATGACA TTGTCGTGAA GGTGCTGAAG GTTCGAGACT    60

GGAGTACAAG GAAGAGCAGG GACTTCAATG AAGAGTGTCC CCGGCTCAGG ATTTTTCGCA   120

TCCAAATGTG CTCCCAGTGC TAGGTGCCTG CC                                 152
```

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 152 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Rheumatoid Synovium
(B) CLONE: 191283

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:36:

```
CAACTACAGT GAACCTAAAA TGCCTCTAAT ACCTTTGCAA TTATCTTTAA GAGGATATCT    60

TATGAGTGAA ATTAACTTGT GCAACTACTT TCCTATTCAC TTTTTTACAG AGACTTAAAA   120

CCAGAGAATA TTTCTAGATT CACAGGGACA CT                                 152
```

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 199 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Rheumatoid Synovium
(B) CLONE: 192268

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:37:

```
AGTGGACTGC AGTAAGCAGA GCTTCCTGAC CGAGGTGGAG CAGCTGTCCA GGTTTCGTCA    60

CCCAAACATT GTGGACTTTC TGGCTACTGT GCTCAGAACG GCTTCTACTG CCTGGTGTAC   120

GGCTTCCTGC CCAACGGCTC CCTGGAGGAC CGTTCCACTG CCAGACCCAG GCCTGCCCAC   180

CTCTCTCCTG GCCTCAGCG                                                199
```

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 189 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

-continued

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Stomach
(B) CLONE: 214915

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:38:

```
AGAAGATCCA GTACCTGGTG TATCAATGCT CAAAGGCCTT AAGTACATCC ACTCTCTGGG     60
GTCGTGCACA GGGACCTGAA GCCAGGCAAC CTGGCTGTGA ATAGGACTGT AACTGAAGAT    120
TCTGGATTTT GGGCTGGCGC GACATGCAGA CGCCGAGATG ACTGGCTACG TGGTGACCCG    180
CTGGTACCT                                                            189
```

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 167 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Pancreas
(B) CLONE: 223163

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:39:

```
CTTGCTCTTC TGACAGGATG AGAGTTATTA TAAGCAAATC CTACCTAGAG GCTTTTAACT     60
CTAATGGGAA TAACTTGCAA CTAAAAGACC CAACTTGCAG ACCAAAATTA TCAAATGTTG    120
TGGATTTTCT GTCCCTCTTA ATGGATGTGG TACAATCAGA AAGGTAG                  167
```

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 197 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Small Intestine
(B) CLONE: 237002

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:40:

```
CCCAAACCTG CCCAGCCAGC CCTGAAAATG CAAGTTTTGT ACGATTTTGA AGCTAGGAAC     60
CCACGGGAAC TGACTGTGGT CCAGGGAGAG AAGCTGGAGG TTTGGACCAC AGCAAGCGGT    120
GGTGGCTGGT GAAGAATAGG CGGGACGGAG CGGCTACATT CCAAGCAACA TCTGGGCCCC    180
TACAGCCGGG GACCCCG                                                   197
```

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 207 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Hippocampus
(B) CLONE: 239990

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:41:

-continued

```
CCAAGATGCT GGAGGAACTC AAGCCGAGAC TTGTACCAAG GAGAGATGAG CAGGAAGGAG    60
GCAGAGGGCT CTGAGAAAGA CGGGACTTCC TGGTCAGGAA GAGCACCACC AACCCGGGCT   120
CCTTTTCCTC ACGGGCATGC ACAATGGCCA GGCAAGCACC TGCTGCTCTT GGACCCAGAA   180
GGCACGTCCG GACAAAGGCA GAGTCTT                                      207
```

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 195 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Hippocampus
(B) CLONE: 240142

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:42:

```
GTCACCGGAG AGGATCCATG AGAACGGCTA CAACTTCAAG TCCGACATCT GGTCCTTGGG    60
CTGTCTGCTG TACGAGATGG CAGCCCTCCA GAGCCCCTTC TATGGAGATA AGATGAATCT   120
TTCTCCCTGT GCCAGAAGAT CGAGCAGTGT GACTACCCCC CACTCCCCGG GGAGCACTAC   180
TCCGAGAAGT TACGT                                                   195
```

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 213 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Testes
(B) CLONE: 275781

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:43:

```
CTCGTCTATT CGGCACGAGT TTCATTGTCG AAGGAAATAT AAACTGTCTG GAAGATCTGG    60
TGTAGCTCCT TCGAGACATC TTTGGCGATC AGCATCACCA ACGGTAAGAA GTGTAGTAAG   120
CCAGATCTCA GGGCCAGGCA TCCCCAGTTG CTGTACAAGA GCAGGCTTTC AAGATGCTTC   180
AAGGTCCCTG TCCATCAATA TGCTACACAT TTG                               213
```

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 425 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Eosinophils
(B) CLONE: 285465

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:44:

```
AAATACTTGA AGGAGTTTAT TATCTACATC AGAATAACAT TGTACACCTT GATTTAAAGC    60
CACAGAATAT ATTACTGAGC AGCATATACC CTCTCGGGGA CATTAAAATA GTAGATTTTG   120
GAATGTCTCG AAAAATAGGG CATGCGTGTG AACTTCGGGA AATCATGGGA ACACCAGAAT   180
```

-continued

```
ATTTAGCTCC AGAAATCCTG AACTATGATC CCATTACCAC AGCAACAGAT ATGTGGAATA    240
TTGGTATAAT AGCATATATG TTGTTAACTC ACACATCACC ATTTGTGGGA GAAGATAATC    300
AAGAAACATA CCTCAATATC TCTCAAGTTA ATGTAGATTA TTCGGAAGGA ACTTTTTCAT    360
CAGTTTCACA GCTGGCACAG ACTTTATTCA GAGCTTTTAG TAAAATCAGA GGAAAGGCCC    420
ACAGC                                                                425
```

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1851 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(v i i) IMMEDIATE SOURCE:
(A) LIBRARY: Stomach
(B) CLONE: 214915E

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:45:

```
GCCCGTTGGG CCGCGAACGC AGCCGCCACG CCGGGGCCGC CGAGATCGGG TGCCCGGGAT     60
GAGCCTCATC CGGAAAAAGG GCTTCTACAA GCAGGACGTC AACAAGACCG CCTGGGAGCT    120
GCCCAAGACC TACGTGTCCC CGACGCACGT CGGCAGCGGG GCCTATGGCT CCGTGTGCTC    180
GGCCATCGAC AAGCGGTCAG GGGAGAAGGT GGCCATCAAG AAGCTGAGCC GACCCTTTCA    240
GTCCGAGATC TTCGCCAAGC GCGCCTACCG GGAGCTGCTG TTGCTGAAGC ACATGCAGCA    300
TGAGAACGTC ATTGGGCTCC TGGATGTCTT CACCCCAGCC TCCTCCCTGG AACTTCTATG    360
ACTTCTACCT GGTGATGCCC TTCATGCAGA CGGATCTGCA GAAGATCATG GGGATGGAGT    420
TCAGTGAGGA GAAGATCCAG TACCTGGTGT ATCAGATGCT CAAAGGCCTT AAGTACATCC    480
ACTCTGCTGG GGTCGTGCAC AGGGACCTGA AGCCAGGCAA CCTGGCTGTG AATGAGGACT    540
GTGAACTGAA GATTCTGGAT TTGGGGCTGG CGCGACATGC AGACGCCGAG ATGACTGGCT    600
ACGTGGTGAC CCGCTGGTAC CGAGCCCCCG AGGTGATCCT CAGCTGGATG CACTACAACC    660
AGACAGTGGA CATCTGGTCT GTGGGCTGTA TCATGGCAGA GATGCTGACA GGGAAAACTC    720
TGTTCAAGGG GAAAGATTAC CTGGACCAGC TGACCCAGAT CCTGAAAGTG ACCGGGGTGC    780
CTGGCACGGA GTTTGTGCAG AAGCTGAACG ACAAAGCGGC CAAATCCTAC ATCCAGTCCC    840
TGCCACAGAC CCCCAGGAAG GATTTCACTC AGCTGTTCCC ACGGGCCAGC CCCCAGCCTG    900
CGGACCTGCT GGAGAAGATG CTGGAGCTAG ACGTGGACAA GCGCCTGACG GCCGCGCAGG    960
CCCTCACCCA TCCCTTCTTT GAACCCTTCC GGGACCCTGA GGAAGAGACG GAGGCCCAGC   1020
AGCCGTTTGA TGATTCCTTA GAACACGAGA AACTCACAGT GGATGAATGG AAGCAGCACA   1080
TCTACAAGGA GATTGTGAAC TTCAGCCCCA TTGCCCGGAA GGACTCACGG CGCCGGAGTG   1140
GCATGAAGCT GTAGGGACTC ATCTTGCATG GCACCGCCGG CCAGACACTG CCCAAGGACC   1200
AGTATTTGTC ACTACCAAAC TCAGCCCTTC TTGGAATACA GCCTTTCAAG CAGAGGACAG   1260
AAGGGTCCTT CTCCTTATGT GGGAAATGGG CCTAGTAGAT GCAGAATTCA AAGATGTCGG   1320
TTGGGAGAAA CTAGCTCTGA TCCTAACAGG CCACGTTAAA CTGCCCATCT GGAGAATCGC   1380
CTGCAGGTGG GGCCCTTTCC TTCCCGCCAG AGTGGGGCTG AGTGGGCGCT GAGCCAGGCC   1440
GGGGGCCTAT GGCAGTGATG CTGTGTTGGT TTCCTAGGGA TGCTCTAACG AATTACCACA   1500
AACCTGGTGG ATTGAAACAG CAGAACTTGA TTCCCTTACA GTTCTGGAGG CTGGAAATCT   1560
```

-continued

```
GGGATGGAGG TGTTGGCAGG GCTGTGGTCC CTTTGAAGGC TCTGGGGAAG AATCCTTCCT   1620
TGGCTCTTTT TAGCTTGTGG CGGCAGTGGG CAGTCCGTGG CATTCCCCAG CTTATTGCTG   1680
CATCACTCCA GTCTCTGTCT CTTCTGTTCT CTCCTCTTTT AACAACAGTC ATTGGATTTA   1740
GGGCCCACCC TAATCCTGTG TGATCTTATC TTGATCCTTA TTAATTAAAC CTGCAAATAC   1800
TCTAGTTCCA AATAAAGTCA CATTCTCAGG TAAAAAAAAA AAAAAAAAAA A            1851
```

We claim:

1. A purified polynucleotide having a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO: 4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, SEQ ID NO:35, SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, and SEQ ID NO:44.

2. An expression vector comprising the polynucleotide of claim 1.

3. A host cell transformed with the expression vector of claim 2.

4. A method for producing and purifying a polypeptide, said method comprising the steps of:

a) culturing the host cell of claim 3 under conditions suitable for the expression of the peptide; and

b) recovering the polypeptide from the host cell culture.

* * * * *

66

US005654173A

United States Patent [19]

Jacobs et al.

[11] Patent Number: 5,654,173

[45] Date of Patent: Aug. 5, 1997

[54] SECRETED PROTEINS AND POLYNUCLEOTIDES ENCODING THEM

[75] Inventors: Kenneth Jacobs, Newton; John M. McCoy, Reading; Edward R. LaVallie, Tewksbury; Lisa A. Racie; David Merberg, both of Acton; Maurice Treacy, Chestnut Hill; Vikki Spaulding, Billerica, all of Mass.

[73] Assignee: Genetics Institute, Inc., Cambridge, Mass.

[21] Appl. No.: 702,080

[22] Filed: Aug. 23, 1996

[51] Int. Cl.^{6} C12P 21/02; C12N 1/21; C12N 5/10; C07H 21/04

[52] U.S. Cl. 435/69.1; 435/252.3; 435/326; 536/23.5

[58] Field of Search 435/69.1, 326, 435/252.3; 536/23.5

[56] References Cited

PUBLICATIONS

Hunt. Human DNA sequence from cosmid L190B4, Huntington's Disease Region, chromosome 4p 16.3. Direct submission to GenBank, Accession No. Z68276. Dec. 19, 1995.

Sambrook et al. Molecular Cloning: A Laboratory Manual. 2d ed. CSHL Press, Cold Spring Harbor, NY. Chapters 9 and 11. 1989.

Primary Examiner—Vasu S. Jagannathan
Assistant Examiner—Brian Lathrop
Attorney, Agent, or Firm—Scott A. Brown; Thomas J. DesRosier

[57] ABSTRACT

Novel polynucleotides and the proteins encoded thereby are disclosed.

14 Claims, No Drawings

SECRETED PROTEINS AND POLYNUCLEOTIDES ENCODING THEM

FIELD OF THE INVENTION

The present invention provides novel polynucleotides and proteins encoded by such polynucleotides, along with therapeutic, diagnostic and research utilities for these polynucleotides and proteins.

BACKGROUND OF THE INVENTION

Technology aimed at the discovery of protein factors (including e.g., cytokines, such as lymphokines, interferons, CSFs and interleukins) has matured rapidly over the past decade. The now routine hybridization cloning and expression cloning techniques clone novel polynucleotides "directly" in the sense that they rely on information directly related to the discovered protein (i.e., partial DNA/amino acid sequence of the protein in the case of hybridization cloning; activity of the protein in the case of expression cloning). More recent "indirect" cloning techniques such as signal sequence cloning, which isolates DNA sequences based on the presence of a now well-recognized secretory leader sequence motif, as well as various PCR-based or low stringency hybridization cloning techniques, have advanced the state of the art by making available large numbers of DNA/amino acid sequences for proteins that are known to have biological activity by virtue of their secreted nature in the case of leader sequence cloning, or by virtue of the cell or tissue source in the case of PCR-based techniques. It is to these proteins and the polynucleotides encoding them that the present invention is directed.

SUMMARY OF THE INVENTION

In one embodiment, the present invention provides a composition comprising an isolated polynucleotide selected from the group consisting of:

(a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1;

(b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 247 to nucleotide 432;

(c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 328 to nucleotide 432;

(d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146;

(e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146;

(f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146;

(g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146;

(h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:2;

(i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:2 having biological activity;

(j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above;

(k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:1 from nucleotide 247 to nucleotide 432; the nucleotide sequence of SEQ ID NO:1 from nucleotide 328 to nucleotide 432; the nucleotide sequence of the full length protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146; or the nucleotide sequence of the mature protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146.

Other embodiments provide the gene corresponding to the cDNA sequence of SEQ ID NO:1 or SEQ ID NO:3.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

(a) the amino acid sequence of SEQ ID NO:2;

(b) fragments of the amino acid sequence of SEQ ID NO:2; and

(c) the amino acid sequence encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146; the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:2.

In one embodiment, the present invention provides a composition comprising an isolated polynucleotide selected from the group consisting of:

(a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:4;

(b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:4 from nucleotide 316 to nucleotide 501;

(c) a polynucleotide comprising the nucleotide sequence of the full length protein, coding sequence of clone BR533_4 deposited under accession number ATCC 98146;

(d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BR533_4 deposited under accession number ATCC 98146;

(e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BR533_4 deposited under accession number ATCC 98146;

(f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BR533_4 deposited under accession number ATCC 98146;

(g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:5;

(h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:5 having biological activity;

(i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(d) above;

(j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:4 from nucleotide 316 to nucleotide 501; the nucleotide sequence of the full length protein coding sequence of clone BR533_4 deposited under accession number ATCC 98146; or the nucleotide sequence of the mature protein coding sequence of clone BR533_4 deposited under accession number ATCC 98146. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone BR533_4 deposited under accession number ATCC 98146.

Other embodiments provide the gene corresponding to the cDNA sequence of SEQ ID NO:4 or SEQ ID NO:6.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

(a) the amino acid sequence of SEQ ID NO:5;

(b) fragments of the amino acid sequence of SEQ ID NO:5; and

(c) the amino acid sequence encoded by the cDNA insert of clone

BR533_4 deposited under accession number ATCC 98146; the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:5.

In one embodiment, the present invention provides a composition comprising an isolated polynucleotide selected from the group consisting of:

(a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:7;

(b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:7 from nucleotide 113 to nucleotide 433;

(c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone CC288_9 deposited under accession number ATCC 98146;

(d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone CC288_9 deposited under accession number ATCC 98146;

(e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone CC288_9 deposited under accession number ATCC 98146;

(f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone CC288_9 deposited under accession number ATCC 98146;

(g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:8;

(h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:8 having biological activity;

(i) a polynucleotide which is an allelic variant of a polynucleotide of (a)–(d) above;

(j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ D NO:7 from nucleotide 113 to nucleotide 433; the nucleotide sequence of the full length protein coding sequence of clone CC288_9 deposited under accession number ATCC 98164; or the nucleotide sequence of the mature protein coding sequence of clone CC288_9 deposited under accession number ATCC 98146. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone CC288_9 deposited under accession number ATCC 98146. In yet other preferred embodiments, the present invention provides a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:8 from amino acid 1 to amino acid 77.

Other embodiments provide the gene corresponding to the cDNA sequence of SEQ ID NO:7.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

(a) the amino acid sequence of SEQ ID NO:8;

the amino acid sequence of SEQ ID NO:8 from amino acid 1 to amino acid 77;

(c) fragments of the amino acid sequence of SEQ ID NO:8; and

(d) the amino acid sequence encoded by the cDNA insert of clone

CC288_9 deposited under accession number ATCC 98146; the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:8 or the amino acid sequence of SEQ ID NO:8 from amino acid 1 to amino acid 77.

In certain preferred embodiments, the polynucleotide is operably linked to an expression control sequence. The invention also provides a host cell, including bacterial, yeast, insect and mammalian cells, transformed with such polynucleotide compositions.

Processes are also provided for producing a protein, which comprise:

(a) growing a culture of the host cell transformed with such polynucleotide compositions in a suitable culture medium; and

(b) purifying the protein from the culture.

The protein produced according to such methods is also provided by the present invention. Preferred embodiments include those in which the protein produced by such process is a mature form of the protein.

Protein compositions of the present invention may further comprise a pharmaceutically acceptable carrier. Compositions comprising an antibody which specifically reacts with such protein are also provided by the present invention.

Methods are also provided for preventing, treating or ameliorating a medical condition which comprises administering to a mammalian subject a therapeutically effective amount of a composition comprising a protein of the present invention and a pharmaceutically acceptable carrier.

DETAILED DESCRIPTION

ISOLATED PROTEINS AND POLYNUCLEOTIDES

Nucleotide and amino acid sequences are reported below for each clone and protein disclosed in the present application. In some instances the sequences are preliminary and may include some incorrect or ambiguous bases or amino acids. The actual nucleotide sequence of each clone can readily be determined by sequencing of the deposited clone in accordance with known methods. The predicted amino acid sequence (both full length and mature) can then be determined from such nucleotide sequence. The amino acid sequence of the protein encoded by a particular clone can also be determined by expression of the clone in a suitable host cell, collecting the protein and determining its sequence.

For each disclosed protein applicants have identified what they have determined to be the reading frame best identifiable with sequence information available at the time of filing. Because of the partial ambiguity in reported sequence information, reported protein sequences include "Xaa" designators. These "Xaa" designators indicate either (1) a residue which cannot be identified because of nucleotide sequence ambiguity or (2) a stop codon in the determined nucleotide sequence where applicants believe one should not exist (if the nucleotide sequence were determined more accurately).

As used herein a "secreted" protein is one which, when expressed in a suitable host cell, is transported across or through a membrane, including transport as a result of signal sequences in its amino acid sequence. "Secreted" proteins include without limitation proteins secreted wholly (e.g., soluble proteins) or partially (e.g., receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins which are transported across the membrane of the endoplasmic reticulum.

Clone "BD372 5"

A polynucleotide of the present invention has been identified as clone "BD372_5". BD372_5 was isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. BD372_5 is a full-length clone, including the entire coding sequence of a secreted protein (also referred to herein as "BD372_5 protein").

The nucleotide sequence of the 5' portion of BD372_5 as presently determined is reported in SEQ ID NO:1. What applicants presently believe is the proper reading frame for the coding region is indicated in SEQ ID NO:2. The predicted acid sequence of the BD372_5 protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:2. Amino acids 1 to 27 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 28. Additional nucleotide sequence from the 3' portion of BD372_5, including the polyA tail, is reported in SEQ ID NO:3.

The EcoRI/NotI restriction fragment obtainable from the deposit containing clone BD372_5 should be approximately 2300 bp.

The nucleotide sequence disclosed herein for BD372_5 was searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. BD372_5 demonstrated at least some identity with ESTs identified as "yc90f12.s 1 Homo sapiens cDNA clone 23278 3'" (R39276, BlastN) and "EST05537 Homo sapiens cDNA clone HFBEM26" (T07647, Fasta). Based upon identity, BD372_5 proteins and each identical protein or peptide may share at least some activity.

Clone "BR533 4"

A polynucleotide of the present invention has been identified as clone "BR533_4". BR533_4 was isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. BR533_4 is a full-length clone, including the entire coding sequence of a secreted protein (also referred to herein as "BR533_4 protein").

The nucleotide sequence of the 5' portion of BR533_4 as presently determined is reported in SEQ ID NO:4. What applicants presently believe is the proper reading frame for the coding region is indicated in SEQ ID NO:5. The predicted acid sequence of the BR533_4 protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:5. Additional nucleotide sequence from the 3' portion of BR533_4, including the polyA tail, is reported in SEQ ID NO:6.

The EcoRI/NotI restriction fragment obtainable from the deposit containing clone BR533_4 should be approximately 2850 bp.

The nucleotide sequence disclosed herein for BR533_4 was searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. BR533_4 demonstrated at least some homology with murine semaphorin E (X85994, BlastN). BR533_4 also shows at least some identity with an EST identified as "yy80d10.s 1 Homo sapiens cDNA clone 279859 3'" (N38844, BlastN). Based upon homology, BR533_4 proteins and each homologous protein or peptide may share at least some activity.

Clone "CC288 9"

A polynucleotide of the present invention has been identified as clone "CC288_9". CC288_9 was isolated from a human adult brain cDNA library using methods which are selective for cDNAs encoding secreted proteins. CC288_9 is a full-length clone, including the entire coding sequence of a secreted protein (also referred to herein as "CC288_9 protein").

The nucleotide sequence of CC288_9 as presently determined is reported in SEQ ID NO:7. What applicants presently believe to be the proper reading frame and the predicted amino acid sequence of the CC288_9 protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:8.

The nucleotide sequence disclosed herein for CC288_9 was searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. No hits were found in the database.

Deposit of Clones

Clones BD372_5, BR533_4 and CC288_9 were deposited on Aug. 22, 1996 with the American Type Culture Collection under accession number ATCC 98146, from which each clone comprising a particular polynucleotide is obtainable. Each clone has been transfected into separate bacterial cells (*E. coli*) in this composite deposit. Each clone can be removed from the vector in which it was deposited by performing an EcoRI/NotI digestion (5' cite, EcoRI; 3' cite, NotI) to produce the appropriately sized fragment for such clone (approximate clone size fragment are identified below). Bacterial cells containing a particular clone can be obtained from the composite deposit as follows:

An oligonucleotide probe or probes should be designed to the sequence that is known for that particular clone. This sequence can be derived from the sequences provided herein, or from a combination of those sequences. The sequence of the oligonucleotide probe that was used to isolate each full-length clone is identified below, and should be most reliable in isolating the clone of interest.

Clone	Probe Sequence
BD372_5	SEQ ID NO: 9
BR533_4	SEQ ID NO: 10
CC288_9	SEQ ID NO: 11

In the sequences listed above which include an N at position 2, that position is occupied in preferred probes/primers by a biotinylated phosphoaramidite residue rather than a nucleotide (such as, for example, that produced by use of biotin phosphoramidite (1-dimethoxytrityloxy-2-(N-biotinyl-4-aminobutyl)-propyl-3-O-(2-cyanoethyl)-(N,N-diisopropyl)-phosphoramadite) (Glen Research, cat. no. 10-1953)).

The design of the oligonucleotide probe should preferably follow these parameters:

(a) It should be designed to an area of the sequence which has the fewest ambiguous bases ("N's"), if any;

(b) It should be designed to have a T_m of approx. 80° C. (assuming 2° for each A or T and 4 degrees for each G or C).

The oligonucleotide should preferably be labeled with g-^{32}P ATP (specific activity 6000 Ci/mmole) and T4 polynucleotide kinase using commonly employed techniques for labeling oligonucleotides. Other labeling techniques can also be used. Unincorporated label should preferably be removed by gel filtration chromatography or other established methods. The amount of radioactivity incorporated into the probe should be quantitated by measurement in a scintillation counter. Preferably, specific activity of the resulting probe should be approximately 4e+6 dmp/pmole.

The bacterial culture containing the pool of full-length clones should preferably be thawed and 100 μl of the stock used to inoculate a sterile culture flask containing 25 ml of sterile L-broth containing ampicillin at 100 μg/mL. The culture should preferably be grown to saturation at 37° C., and the saturated culture should preferably be diluted in fresh L-broth. Aliquots of these dilutions should preferably be plated to determine the dilution and volume which will yield approximately 5000 distinct and well-separated colonies on solid bacteriological media containing L-broth containing ampicillin at 100 μg/ml and agar at 1.5% in a 150 mm petri dish when grown overnight at 37° C. Other known methods of obtaining distinct, well-separated colonies can also be employed.

Standard colony hybridization procedures should then be used to transfer the colonies to nitrocellulose filters and lyse, denature and bake them.

The filter is then preferably incubated at 65° C. for 1 hour with gentle agitation in 6× SSC (20× stock is 175.3 g NaCl/liter, 88.2 g Na citrate/liter, adjusted to pH 7.0 with NaOH) containing 0.5% SDS, 100 μg/ml of yeast RNA, and 10 mM EDTA (approximately 10 mL per 150 mm filter). Preferably, the probe is then added to the hybridization mix at a concentration greater than or equal to 1e+6 dpm/mL. The filter is then preferably incubated at 65° C. with gentle agitation overnight. The filter is then preferably washed in 500 mL of 2× SSC/0.5% SDS at room temperature without agitation, preferably followed by 500 mL of 2× SSC/0.1% SDS at room temperature with gentle shaking for 15 minutes. A third wash with 0.1× SSC/0.5% SDS at 65° C. for 30 minutes to 1 hour is optional. The filter is then preferably dried and subjected to autoradiography for sufficient time to visualize the positives on the X-ray film. Other known hybridization methods can also be employed.

The positive colonies are picked, grown in culture, and plasmid DNA isolated using standard procedures. The clones can then be verified by restriction analysis, hybridization analysis, or DNA sequencing.

Fragments of the proteins of the present invention which are capable of exhibiting biological activity are also encompassed by the present invention. Fragments of the protein may be in linear form or they may be cyclized using known methods, for example, as described in H. U. Saragovi, et al., Bio/Technology 10, 773–778 (1992) and in R. S. McDowell, et al., J. Amer. Chem. Soc. 114, 9245–9253 (1992), both of which are incorporated herein by reference. Such fragments may be fused to carrier molecules such as immunoglobulins for many purposes, including increasing the valency of protein binding sites. For example, fragments of the protein may be fused through "linker" sequences to the Fc portion of an immunoglobulin. For a bivalent form of the protein, such a fusion could be to the Fc portion of an IgG molecule. Other immunoglobulin isotypes may also be used to generate such fusions. For example, a protein—IgM fusion would generate a decavalent form of the protein of the invention.

The present invention also provides both full-length and mature forms of the disclosed proteins. The full-length form of the such proteins is identified in the sequence listing by translation of the nucleotide sequence of each disclosed clone. The mature form of such protein may be obtained by expression of the disclosed full-length polynucleotide (preferably those deposited with ATCC) in a suitable mammalian cell or other host cell. The sequence of the mature form of the protein may also be determinable from the amino acid sequence of the full-length form.

The present invention also provides genes corresponding to the cDNA sequences disclosed herein. The corresponding genes can be isolated in accordance with known methods using the sequence information disclosed herein. Such methods include the preparation of probes or primers from the disclosed sequence information for identification and/or amplification of genes in appropriate genomic libraries or other sources of genomic materials.

Where the protein of the present invention is membrane-bound (e.g., is a receptor), the present invention also provides for soluble forms of such protein. In such forms part or all of the intracellular and transmembrane domains of the protein are deleted such that the protein is fully secreted from the cell in which it is expressed. The intracellular and transmembrane domains of proteins of the invention can be identified in accordance with known techniques for determination of such domains from sequence information.

Species homologs of the disclosed polynucleotides and proteins are also provided by the present invention. Species homologs may be isolated and identified by making suitable probes or primers from the sequences provided herein and screening a suitable nucleic acid source from the desired species.

The invention also encompasses allelic variants of the disclosed polynucleotides or proteins; that is, naturally-occurring alternative forms of the isolated polynucleotide which also encode proteins which are identical, homologous or related to that encoded by the polynucleotides.

The isolated polynucleotide of the invention may be operably linked to an expression control sequence such as the pMT2 or pED expression vectors disclosed in Kaufman et al., Nucleic Acids Res. 19, 4485–4490 (1991), in order to produce the protein recombinantly. Many suitable expression control sequences are known in the art. General methods of expressing recombinant proteins are also known and are exemplified in R. Kaufman, Methods in Enzymology 185, 537–566 (1990). As defined herein "operably linked" means that the isolated polynucleotide of the invention and an expression control sequence are situated within a vector or cell in such a way that the protein is expressed by a host cell which has been transformed (transfected) with the ligated polynucleotide/expression control sequence.

A number of types of cells may act as suitable host cells for expression of the protein. Mammalian host cells include, for example, monkey COS cells, Chinese Hamster Ovary (CHO) cells, human kidney 293 cells, human epidermal A431 cells, human Colo205 cells, 3T3 cells, CV-1 cells, other transformed primate cell lines, normal diploid cells, cell strains derived from in vitro culture of primary tissue, primary explants, HeLa cells, mouse L cells, BHK, HL-60, U937, HaK or Jurkat cells.

Alternatively, it may be possible to produce the protein in lower eukaryotes such as yeast or in prokaryotes such as bacteria. Potentially suitable yeast strains include *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, Kluyveromyces strains, Candida, or any yeast strain capable of expressing heterologous proteins. Potentially suitable bacterial strains include *Escherichia coli*, *Bacillus subtilis*, *Salmonella typhimurium*, or any bacterial strain capable of expressing heterologous proteins. If the protein is made in yeast or bacteria, it may be necessary to modify the protein produced therein, for example by phosphorylation or glycosylation of the appropriate sites, in order to obtain the functional protein. Such covalent attachments may be accomplished using known chemical or enzymatic methods.

The protein may also be produced by operably linking the isolated polynucleotide of the invention to suitable control sequences in one or more insect expression vectors, and employing an insect expression system. Materials and methods for baculovirus/insect cell expression systems are commercially available in kit form from, e.g., Invitrogen, San Diego, Calif., U.S.A. (the MaxBat® kit), and such methods are well known in the art, as described in Summers and Smith, *Texas Agricultural Experiment Station Bulletin* No. 1555 (1987), incorporated herein by reference. As used herein, an insect cell capable of expressing a polynucleotide of the present invention is "transformed."

The protein of the invention may be prepared by culturing transformed host cells under culture conditions suitable to express the recombinant protein. The resulting expressed protein may then be purified from such culture (i.e., from culture medium or cell extracts) using known purification processes, such as gel filtration and ion exchange chromatography. The purification of the protein may also include an affinity column containing agents which will bind to the protein; one or more column steps over such affinity resins as concanavalin A-agarose, heparin-toyopearl® or Cibacrom blue 3GA Sepharose®; one or more steps involving hydrophobic interaction chromatography using such resins as phenyl ether, butyl ether, or propyl ether; or immunoaffinity chromatography.

Alternatively, the protein of the invention may also be expressed in a form which will facilitate purification. For example, it may be expressed as a fusion protein, such as those of maltose binding protein (MBP), glutathione-S-transferase (GST) or thioredoxin (TRX). Kits for expression and purification of such fusion proteins are commercially available from New England BioLab (Beverly, Mass.), Pharmacia (Piscataway, N.J.) and In Vitrogen, respectively. The protein can also be tagged with an epitope and subsequently purified by using a specific antibody directed to such epitope. One such epitope ("Flag") is commercially available from Kodak (New Haven, Conn.).

Finally, one or more reverse-phase high performance liquid chromatography (RP-HPLC) steps employing hydrophobic RP-HPLC media, e.g., silica gel having pendant methyl or other aliphatic groups, can be employed to further purify the protein. Some or all of the foregoing purification steps, in various combinations, can also be employed to provide a substantially homogeneous isolated recombinant protein. The protein thus purified is substantially free of other mammalian proteins and is defined in accordance with the present invention as an "isolated protein."

The protein of the invention may also be expressed as a product of transgenic animals, e.g., as a component of the milk of transgenic cows, goats, pigs, or sheep which are characterized by somatic or germ cells containing a nucleotide sequence encoding the protein.

The protein may also be produced by known conventional chemical synthesis. Methods for constructing the proteins of the present invention by synthetic means are known to those skilled in the art. The synthetically-constructed protein sequences, by virtue of sharing primary, secondary or tertiary structural and/or conformational characteristics with proteins may possess biological properties in common therewith, including protein activity. Thus, they may be employed as biologically active or immunological substitutes for natural, purified proteins in screening of therapeutic compounds and in immunological processes for the development of antibodies.

The proteins provided herein also include proteins characterized by amino acid sequences similar to those of purified proteins but into which modification are naturally provided or deliberately engineered. For example, modifications in the peptide or DNA sequences can be made by those skilled in the art using known techniques. Modifications of interest in the protein sequences may include the alteration, substitution, replacement, insertion or deletion of a selected amino acid residue in the coding sequence. For example, one or more of the cysteine residues may be deleted or replaced with another amino acid to alter the conformation of the molecule. Techniques for such alteration, substitution, replacement, insertion or deletion are well known to those skilled in the art (see, e.g., U.S. Pat. No. 4,518,584). Preferably, such alteration, substitution, replacement, insertion or deletion retains the desired activity of the protein.

Other fragments and derivatives of the sequences of proteins which would be expected to retain protein activity in whole or in part and may thus be useful for screening or other immunological methodologies may also be easily made by those skilled in the art given the disclosures herein. Such modifications are believed to be encompassed by the present invention.

USES AND BIOLOGICAL ACTIVITY

The polynucleotides and proteins of the present invention are expected to exhibit one or more of the uses or biological activities (including those associated with assays cited herein) identified below. Uses or activities described for proteins of the present invention may be provided by administration or use of such proteins or by administration or use of polynucleotides encoding such proteins (such as, for example, in gene therapies or vectors suitable for introduction of DNA).

Research Uses and Utilities

The polynucleotides provided by the present invention can be used by the research community for various purposes. The polynucleotides can be used to express recombinant protein for analysis, characterization or therapeutic use; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in disease states); as molecular weight markers on Southern gels; as chromosome markers or tags (when labeled) to identify chromosomes or to map related gene positions; to compare with endogenous DNA sequences in patients to identify potential genetic disorders; as probes to hybridize and thus discover novel, related DNA sequences; as a source of information to derive PCR primers for genetic fingerprinting; as a probe to "subtract-out" known sequences in the process of discovering other novel polynucleotides; for selecting and making oligomers for attachment to a "gene chip" or other support, including for examination of expression patterns; to raise anti-protein antibodies using DNA immunization techniques; and as an antigen to raise anti-DNA antibodies or elicit another immune response. Where the polynucleotide encodes a protein which binds or potentially binds to another protein (such as, for example, in a receptor-ligand interaction), the polynucleotide can also be used in interaction trap assays (such as, for example, that described in Gyuris et al., Cell 75:791–803 (1993)) to identify polynucleotides encoding the other protein with which binding occurs or to identify inhibitors of the binding interaction.

The proteins provided by the present invention can similarly be used in assay to determine biological activity, including in a panel of multiple proteins for high-throughput screening; to raise antibodies or to elicit another immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the protein (or its receptor) in biological fluids; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state); and, of course, to isolate correlative receptors or ligands. Where the protein binds or potentially binds to another protein (such as, for example, in a receptor-ligand interaction), the protein can be used to identify the other protein with which binding occurs or to identify inhibitors of the binding interaction. Proteins involved in these binding interactions can also be used to screen for peptide or small molecule inhibitors or agonists of the binding interaction.

Any or all of these research utilities are capable of being developed into reagent grade or kit format for commercialization as research products.

Methods for performing the uses listed above are well known to those skilled in the art. References disclosing such methods include without limitation "Molecular Cloning: A Laboratory Manual", 2d ed., Cold Spring Harbor Laboratory Press, Sambrook, J., E. F. Fritsch and T. Maniatis eds., 1989, and "Methods in Enzymology: Guide to Molecular Cloning Techniques", Academic Press, Berger, S. L. and A. R. Kimmel eds., 1987.

Nutritional Uses

Polynucleotides and proteins of the present invention can also be used as nutritional sources or supplements. Such uses include without limitation use as a protein or amino acid supplement, use as a carbon source, use as a nitrogen source and use as a source of carbohydrate. In such cases the protein or polynucleotide of the invention can be added to the feed of a particular organism or can be administered as a separate solid or liquid preparation, such as in the form of powder, pills, solutions, suspensions or capsules. In the case of microorganisms, the protein or polynucleotide of the invention can be added to the medium in or on which the microorganism is cultured.

Cytokine and Cell Proliferation/Differentiation Activity

A protein of the present invention may exhibit cytokine, cell proliferation (either inducing or inhibiting) or cell differentiation (either inducing or inhibiting) activity or may induce production of other cytokines in certain cell populations. Many protein factors discovered to date, including all known cytokines, have exhibited activity in one or more factor dependent cell proliferation assays, and hence the assays serve as a convenient confirmation of cytokine activity. The activity of a protein of the present invention is evidenced by any one of a number of routine factor dependent cell proliferation assays for cell lines including, without limitation, 32D, DA2, DA1G, T10, B9, B9/11, BaF3, MC9/G, M+(preB M+), 2E8, RB5, DA1, 123, T1165, HT2, CTLL2, TF-1, Mo7e and CMK.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for T-cell or thymocyte proliferation include without limitation those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function 3.1–3.19; Chapter 7, Immunologic studies in Humans); Takai et al., J. Immunol. 137:3494–3500, 1986; Bertagnolli et al., J. Immunol. 145:1706–1712, 1990; Bertagnolli et al., Cellular Immunology 133:327–341, 1991; Bertagnolli, et al., I. Immunol. 149:3778–3783, 1992; Bowman et al., I. Immunol. 152:1756–1761, 1994.

Assays for cytokine production and/or proliferation of spleen cells, lymph node cells or thymocytes include, without limitation, those described in: Polyclonal T cell stimulation, Kruisbeek, A. M. and Shevach, E. M. In *Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 3.12.1–3.12.14, John Wiley and Sons, Toronto. 1994; and Measurement of mouse and human interleukin γ, Schreiber, R. D. In *Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 6.8.1–6.8.8, John Wiley and Sons, Toronto. 1994.

Assays for proliferation and differentiation of hematopoietic and lymphopoietic cells include, without limitation, those described in: Measurement of Human and Murine Interleukin 2 and Interleukin 4, Bottomly, K., Davis, L. S. and Lipsky, P. E. In *Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 6.3.1–6.3.12, John Wiley and Sons, Toronto. 1991; deVries et al., J. Exp. Med. 173:1205–1211, 1991; Moreau et al., Nature 336:690–692, 1988; Greenberger et al., Proc. Natl. Acad. Sci. U.S.A. 80:2931–2938, 1983; Measurement of mouse and human interleukin 6—Nordan, R. In *Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 6.6.1–6.6.5, John Wiley and Sons, Toronto. 1991; Smith et al., Proc. Natl. Aced. Sci. U.S.A. 83:1857–1861, 1986; Measurement of human Interleukin 11—Bennett, F., Giannotti, J., Clark, S. C. and Turner, K. J. *In Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 6.15.1 John Wiley and Sons, Toronto. 1991; Measurement of mouse and human Interleukin 9—Ciarletta, A., Giannotti, J., Clark, S. C. and Turner, K. J. *In Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 6.13.1, John Wiley and Sons, Toronto. 1991.

Assays for T-cell clone responses to antigens (which will identify, among others, proteins that affect APC-T cell interactions as well as direct T-cell effects by measuring proliferation and cytokine production) include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function; Chapter 6, Cytokines and their cellular receptors; Chapter 7, Immunologic studies in Humans); Weinberger et al., Proc. Natl. Acad. Sci. USA 77:6091–6095, 1980; Weinberger et al., Eur. J. Immun. 11:405–411, 1981; Takai et al., J. Immunol. 137:3494–3500, 1986; Takai et al., J. Immunol. 140:508–512, 1988.

Immune Stimulating or Suppressing Activity

A protein of the present invention may also exhibit immune stimulating or immune suppressing activity, including without limitation the activities for which assays are described herein. A protein may be useful in the treatment of various immune deficiencies and disorders (including severe combined immunodeficiency (SCID)), e.g., in regulating (up or down) growth and proliferation of T and/or B lymphocytes, as well as effecting the cytolytic activity of NK cells and other cell populations. These immune deficiencies may be genetic or be caused by vital (e.g., HIV) as well as bacterial or fungal infections, or may result from autoimmune disorders. More specifically, infectious diseases causes by viral, bacterial, fungal or other infection may be treatable using a protein of the present invention, including infections by HIV, hepatitis viruses, herpesviruses, mycobacteria, Leishmania spp., malaria spp. and various fungal infections such as candidiasis. Of course, in this regard, a protein of the present invention may also be useful where a boost to the immune system generally may be desirable, i.e., in the treatment of cancer.

Autoimmune disorders which may be treated using a protein of the present invention include, for example, connective tissue disease, multiple sclerosis, systemic lupus erythematosus, rheumatoid arthritis, autoimmune pulmonary inflammation, Guillain-Barre syndrome, autoimmune thyroiditis, insulin dependent diabetes mellitis, myasthenia gravis, graft-versus-host disease and autoimmune inflammatory eye disease. Such a protein of the present invention may also to be useful in the treatment of allergic reactions and conditions, such as asthma (particularly allergic asthma) or other respiratory problems. Other conditions, in which immune suppression is desired (including, for example, organ transplantation), may also be treatable using a protein of the present invention.

Using the proteins of the invention it may also be possible to immune responses, in a number of ways. Down regulation may be in the form of inhibiting or blocking an immune response already in progress or may involve preventing the induction of an immune response. The functions of activated T cells may be inhibited by suppressing T cell responses or by inducing specific tolerance in T cells, or both. Immunosuppression of T cell responses is generally an active, non-antigen-specific, process which requires continuous exposure of the T cells to the suppressive agent. Tolerance, which involves inducing non-responsiveness or anergy in T cells, is distinguishable from immunosuppression in that it is generally antigen-specific and persists after exposure to the tolerizing agent has ceased. Operationally, tolerance can be demonstrated by the lack of a T cell response upon reexposure to specific antigen in the absence of the tolerizing agent.

Down regulating or preventing one or more antigen functions (including without limitation B lymphocyte antigen functions (such as, for example, B7)), e.g., preventing high level lymphokine synthesis by activated T cells, will be useful in situations of tissue, skin and organ transplantation and in graft-versus-host disease (GVHD). For example, blockage of T cell function should result in reduced tissue destruction in tissue transplantation. Typically, in tissue transplants, rejection of the transplant is initiated through its recognition as foreign by T cells, followed by an immune reaction that destroys the transplant. The administration of a molecule which inhibits or blocks interaction of a B7 lymphocyte antigen with its natural ligand(s) on immune cells (such as a soluble, monomeric form of a peptide having B7-2 activity alone or in conjunction with a monomeric form of a peptide having an activity of another B lymphocyte antigen (e.g., B7-1, B7-3) or blocking antibody), prior to transplantation can lead to the binding of the molecule to the natural ligand(s) on the immune cells without transmitting the corresponding costimulatory signal. Blocking B lymphocyte antigen function in this matter prevents cytokine synthesis by immune cells, such as T cells, and thus acts as an immunosuppressant. Moreover, the lack of costimulation may also be sufficient to anergize the T cells, thereby inducing tolerance in a subject. Induction of long-term tolerance by B lymphocyte antigen-blocking reagents may avoid the necessity of repeated administration of these blocking reagents. To achieve sufficient immunosuppression or tolerance in a subject, it may also be necessary to block the function of a combination of B lymphocyte antigens.

The efficacy of particular blocking reagents in preventing organ transplant rejection or GVHD can be assessed using animal models that are predictive of efficacy in humans. Examples of appropriate systems which can be used include allogeneic cardiac grafts in rats and xenogeneic pancreatic islet cell grafts in mice, both of which have been used to examine the immunosuppressive effects of CTLA4Ig fusion proteins in vivo as described in Lenschow et al., Science 257:789–792 (1992) and Turka et al., Proc. Natl. Acad. Sci USA, 89:11102–11105 (1992). In addition, murine models of GVHD (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 846–847) can be used to determine the effect of blocking B lymphocyte antigen function in vivo on the development of that disease.

Blocking antigen function may also be therapeutically useful for treating autoimmune diseases. Many autoimmune disorders are the result of inappropriate activation of T cells that are reactive against self tissue and which promote the production of cytokines and autoantibodies involved in the pathology of the diseases. Preventing the activation of autoreactive T cells may reduce or eliminate disease symptoms. Administration of reagents which block costimulation of T cells by disrupting receptor:ligand interactions of B lymphocyte antigens can be used to inhibit T cell activation and prevent production of autoantibodies or T cell-derived cytokines which may be involved in the disease process. Additionally, blocking reagents may induce antigen-specific tolerance of autoreactive T cells which could lead to long-term relief from the disease. The efficacy of blocking reagents in preventing or alleviating autoimmune disorders can be determined using a number of well-characterized animal models of human autoimmune diseases. Examples include murine experimental autoimmune encephalitis, systemic lupus erythmatosis in MRL/lpr/lpr mice or NZB hybrid mice, murine autoimmune collagen arthritis, diabetes mellitus in NOD mice and BB rats, and murine experimental myasthenia gravis (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 840–856).

Upregulation of an antigen function (preferably a B lymphocyte antigen function), as a means of up regulating immune responses, may also be useful in therapy. Upregulation of immune responses may be in the form of enhancing an existing immune response or eliciting an initial immune response. For example, enhancing an immune response through stimulating B lymphocyte antigen function may be useful in cases of viral infection. In addition, systemic viral diseases such as influenza, the common cold, and encephalitis might be alleviated by the administration of stimulatory forms of B lymphocyte antigens systemically.

Alternatively, anti-vital immune responses may be enhanced in an infected patient by removing T cells from the patient, costimulating the T cells in vitro with viral antigen-pulsed APCs either expressing a peptide of the present invention or together with a stimulatory form of a soluble peptide of the present invention and reintroducing the in vitro activated T cells into the patient. Another method of enhancing anti-viral immune responses would be to isolate infected cells from a patient, transfect them with a nucleic acid encoding a protein of the present invention as described herein such that the cells express all or a portion of the protein on their surface, and reintroduce the transfected cells into the patient. The infected cells would now be capable of delivering a costimulatory signal to, and thereby activate, T cells in vivo.

In another application, up regulation or enhancement of antigen function (preferably B lymphocyte antigen function) may be useful in the induction of tumor immunity. Tumor cells (e.g., sarcoma, melanoma, lymphoma, leukemia, neuroblastoma, carcinoma) transfected with a nucleic acid encoding at least one peptide of the present invention can be administered to a subject to overcome tumor-specific tolerance in the subject If desired. the tumor cell can be transfected to express a combination of peptides. For example. tumor cells obtained from a patient can be transfected ex vivo with an expression vector directing the expression of a peptide having B7-2-like activity alone, or in conjunction with a peptide having B7-1-like activity and/or B7-3-like activity. The transfected tumor cells are returned to the patient to result in expression of the peptides on the surface of the transfected cell. Alternatively, gene therapy techniques can be used to target a tumor cell for transfection in vivo.

The presence of the peptide of the present invention having the activity of a B lymphocyte antigen(s) on the surface of the tumor cell provides the necessary costimulation signal to T cells to induce a T cell mediated immune response against the transfected tumor cells. In addition, tumor cells which lack MHC class I or MHC class II molecules. or which fail to reexpress sufficient mounts of MHC class I or MHC class II molecules, can be transfected with nucleic acid encoding all or a portion of (e.g., a cytoplasmic-domain truncated portion) of an MHC class I α chain protein and β_2 microglobulin protein or an MHC class II α chain protein and an MHC class II β chain protein to thereby express MHC class I or MHC class II proteins on the cell surface. Expression of the appropriate class I or class II MHC in conjunction with a peptide having the activity of a B lymphocyte antigen (e.g., B7-1, B7-2, B7-3) induces a T cell mediated immune response against the transfected tumor cell. Optionally, a gene encoding an antisense construct which blocks expression of an MHC class II associated protein, such as the invariant chain, can also be cotransfected with a DNA encoding a peptide having the activity of a B lymphocyte antigen to promote presentation of tumor associated antigens and induce tumor specific immunity. Thus, the induction of a T cell mediated immune response in a human subject may be sufficient to overcome tumor-specific tolerance in the subject.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for thymocyte or splenocyte cytotoxicity include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function 3.1–3.19; Chapter 7, Immunologic studies in Humans); Herrmann et al., Proc. Natl. Acad. Sci. USA 78:2488–2492, 1981; Herrmann et al., J. Immunol. 128:1968–1974, 1982; Handa et al., J. Immunol. 135:1564–1572, 1985; Takai et al., I. Immunol. 137:3494–3500, 1986; Takai et al., J. Immunol. 140:508–512, 1988; Herrmann et al., Proc. Natl. Acad. Sci. USA 78:2488–2492, 1981; Herrmann et al., J. Immunol. 128:1968–1974, 1982; Handa et al., J. Immunol. 135:1564–1572, 1985; Takai et al., J. Immunol. 137:3494–3500, 1986; Bowmanet al., J. Virology 61:1992–1998; Takai et al., J. Immunol. 140:508–512, 1988; Bertagnolli et al., Cellular Immunology 133:327–341, 1991; Brown et al., J. Immunol. 153:3079–3092, 1994.

Assays for T-cell-dependent immunoglobulin responses and isotype switching (which will identify, among others, proteins that modulate T-cell dependent antibody responses and that affect Th1/Th2 profiles) include, without limitation, those described in: Maliszewski, J. Immunol. 144:3028–3033, 1990; and Assays for B cell function: In vitro antibody production, Mond, J. J. and Brunswick, M. In *Current Protocols in Immunology.* J. E. e.a. Coligan eds. Vol 1 pp. 3.8.1–3.8.16, John Wiley and Sons, Toronto. 1994.

Mixed lymphocyte reaction (MLR) assays (which will identify, among others, proteins that generate predominantly Th1 and CTL responses) include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function 3.1–3.19; Chapter 7, Immunologic studies in Humans); Takai et al., J. Immunol. 137:3494–3500, 1986; Takai et al., J. Immunol. 140:508–512, 1988; Bertagnolli et al., J. Immunol. 149:3778–3783, 1992.

Dendritic cell-dependent assays (which will identify, among others, proteins expressed by dendritic cells that activate naive T-cells) include, without limitation, those described in: Guery et al., J. Immunol. 134:536–544, 1995; Inaba et al., Journal of Experimental Medicine 173:549–559, 1991; Macatonia et al., Journal of Immunology 154:5071–5079, 1995; Porgador et al., Journal of Experimental Medicine 182:255–260, 1995; Nair et al., Journal of Virology 67:4062–4069, 1993; Huang et al., Science 264:961–965, 1994; Macatonia et al., Journal of Experimental Medicine 169:1255–1264, 1989; Bhardwaj et al., Journal of Clinical Investigation 94:797–807, 1994; and Inaba et al., Journal of Experimental Medicine 172:631–640, 1990.

Assays for lymphocyte survival/apoptosis (which will identify, among others, proteins that prevent apoptosis after superantigen induction and proteins that regulate lymphocyte homeostasis) include, without limitation, those described in: Darzynkiewicz et al., Cytometry 13:795–808, 1992; Gorczyca et al., Leukemia 7:659–670, 1993; Gorczyca et al., Cancer Research 53:1945–1951, 1993; Itoh et al., Cell 66:233–243, 1991; Zacharchuk, Journal of Immunology 145:4037–4045, 1990; Zamai et al., Cytometry 14:891–897, 1993; Gorczyca et al., International Journal of Oncology 1:639–648, 1992.

Assays for proteins that influence early steps of T-cell commitment and development include, without limitation, those described in: Antica et al., Blood 84:111–117, 1994; Fine et al., Cellular Immunology 155:111–122, 1994; Galy et al., Blood 85:2770–2778, 1995; Toki et al., Proc. Nat. Acad Sci. USA 88:7548–7551, 1991.

Hematopoiesis Regulating Activity

A protein of the present invention may be useful in regulation of hematopoiesis and, consequently, in the treatment of myeloid or lymphoid cell deficiencies. Even marginal biological activity in support of colony forming cells or of factor-dependent cell lines indicates involvement in regulating hematopoiesis, e.g. in supporting the growth and proliferation of erythroid progenitor cells alone or in combination with other cytokines, thereby indicating utility, for example, in treating various anemias or for use in conjunction with irradiation/chemotherapy to stimulate the production of erythroid precursors and/or erythroid cells; in supporting the growth and proliferation of myeloid cells such as granulocytes and monocytes/macrophages (i.e., traditional CSF activity) useful, for example, in conjunction with chemotherapy to prevent or treat consequent myelosuppression; in supporting the growth and proliferation of megakaryocytes and consequently of platelets thereby allowing prevention or treatment of various platelet disorders such as thrombocytopenia, and generally for use in place of or complimentary to platelet transfusions; and/or in supporting the growth and proliferation of hematopoietic stem cells which are capable of maturing to any and all of the above-mentioned hematopoietic cells and therefore find therapeutic utility in various stem cell disorders (such as those usually treated with transplantation, including, without limitation, aplastic anemia and paroxysmal nocturnal hemoglobinuria), as well as in repopulating the stem cell compartment post irradiation/chemotherapy, either in-vivo or ex-vivo (i.e., in conjunction with bone marrow transplantation or with peripheral progenitor cell transplantation (homologous or heterologous)) as normal cells or genetically manipulated for gene therapy.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for proliferation and differentiation of various hematopoietic lines are cited above.

Assays for embryonic stem cell differentiation (which will identify, among others, proteins that influence embryonic differentiation hematopoiesis) include, without limitation, those described in: Johansson et al. Cellular Biology 15:141–151, 1995; Keller et al., Molecular and Cellular Biology 13:473–486, 1993; McClanahan et al., Blood 81:2903–2915, 1993.

Assays for stem cell survival and differentiation (which will identify, among others, proteins that regulate lymphohematopoiesis) include, without limitation, those described in: Methylcellulose colony forming assays, Freshney, M. G. In *Culture of Hematopoietic Cells.* R. I. Freshney, et al. eds. Vol pp. 265–268, Wiley-Liss, Inc., New York, N.Y. 1994; Hirayama et al., Proc. Natl. Acad. Sci. USA 89:5907–5911, 1992; Primitive hematopoietic colony forming cells with high proliferative potential, McNiece, I. K. and Briddell, R. A. In *Culture of Hematopoietic Cells.* R. I. Freshney, et al. eds. Vol pp. 23–39, Wiley-Liss, Inc., New York, N.Y. 1994; Neben et al., Experimental Hematology 22:353–359, 1994; Cobblestone area forming cell assay, Ploemacher, R. E. In *Culture of Hematopoietic Cells.* R. I. Freshney, et al. eds. Vol pp. 1–21, Wiley-Liss, Inc., New York, N.Y. 1994; Long term bone marrow cultures in the presence of stromal cells, Spooncer, E., Dexter, M. and Allen, T. In *Culture of Hematopoietic Cells.* R. I. Freshney, et al. eds. Vol pp. 163–179, Wiley-Liss, Inc., New York, N.Y. 1994; Long term culture initiating cell assay, Sutherland, H. J. In *Culture of Hematopoietic Cells.* R. I. Freshney, et al. eds. Vol pp. 139–162, Wiley-Liss, Inc., New York, N.Y. 1994.

Tissue Growth Activity

A protein of the present invention also may have utility in compositions used for bone, cartilage, tendon, ligament and/or nerve tissue growth or regeneration, as well as for wound healing and tissue repair and replacement, and in the treatment of burns, incisions and ulcers.

A protein of the present invention, which induces cartilage and/or bone growth in circumstances where bone is not normally formed, has application in the healing of bone fractures and cartilage damage or defects in humans and other animals. Such a preparation employing a protein of the invention may have prophylactic use in closed as well as open fracture reduction and also in the improved fixation of artificial joints. De novo bone formation induced by an osteogenic agent contributes to the repair of congenital, trauma induced, or oncologic resection induced craniofacial defects, and also is useful in cosmetic plastic surgery.

A protein of this invention may also be used in the treatment of periodontal disease, and in other tooth repair processes. Such agents may provide an environment to attract bone-forming cells, stimulate growth of bone-forming cells or induce differentiation of progenitors of bone-forming cells. A protein of the invention may also be useful in the treatment of osteoporosis or osteoarthritis, such as through stimulation of bone and/or cartilage repair or by blocking inflammation or processes of tissue destruction (collagenase activity, osteoclast activity, etc.) mediated by inflammatory processes.

Another category of tissue regeneration activity that may be attributable to the protein of the present invention is tendon/ligament formation. A protein of the present invention, which induces tendon/ligament-like tissue or other tissue formation in circumstances where such tissue is not normally formed, has application in the healing of tendon or ligament tears, deformities and other tendon or ligament defects in humans and other animals. Such a preparation employing a tendon/ligament-like tissue inducing protein may have prophylactic use in preventing damage to tendon or ligament tissue, as well as use in the improved fixation of tendon or ligament to bone or other tissues, and in repairing defects to tendon or ligament tissue. De novo tendon/ligament-like tissue formation induced by a composition of the present invention contributes to the repair of congenital, trauma induced, or other tendon or ligament defects of other origin, and is also useful in cosmetic plastic surgery for attachment or repair of tendons or ligaments. The compositions of the present invention may provide environment to attract tendon- or ligament-forming cells, stimulate growth of tendon- or ligament-forming cells, induce differentiation of progenitors of tendon- or ligament-forming cells, or induce growth of tendon/ligament cells or progenitors ex vivo for return in vivo to effect tissue repair. The compositions of the invention may also be useful in the treatment of tendinitis, carpal tunnel syndrome and other tendon or ligament defects. The compositions may also include an appropriate matrix and/or sequestering agent as a carrier as is well known in the art.

The protein of the present invention may also be useful for proliferation of neural cells and for regeneration of nerve and brain tissue, i.e. for the treatment of central and peripheral nervous system diseases and neuropathies, as well as mechanical and traumatic disorders, which involve degeneration, death or trauma to neural cells or nerve tissue. More specifically, a protein may be used in the treatment of diseases of the peripheral nervous system, such as peripheral nerve injuries, peripheral neuropathy and localized neuropathies, and central nervous system diseases, such as Alzheimer's, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, and Shy-Drager syndrome. Further conditions which may be treated in accordance with the present invention include mechanical and traumatic disorders, such as spinal cord disorders, head trauma and cerebrovascular diseases such as stroke. Peripheral neuropathies resulting from chemotherapy or other medical therapies may also be treatable using a protein of the invention.

Proteins of the invention may also be useful to promote better or faster closure of non-healing wounds, including without limitation pressure ulcers, ulcers associated with vascular insufficiency, surgical and traumatic wounds, and the like.

It is expected that a protein of the present invention may also exhibit activity for generation or regeneration of other tissues, such as organs (including, for example, pancreas, liver, intestine, kidney, skin, endothelium), muscle (smooth, skeletal or cardiac) and vascular (including vascular endothelium) tissue, or for promoting the growth of cells comprising such tissues. Part of the desired effects may be by inhibition or modulation of fibrotic scarring to allow normal tissue to regenerate. A protein of the invention may also exhibit angiogenic activity.

A protein of the present invention may also be useful for gut protection or regeneration and treatment of lung or liver fibrosis, reperfusion injury in various tissues, and conditions resulting from systemic cytokine damage.

A protein of the present invention may also be useful for promoting or inhibiting differentiation of tissues described above from precursor tissues or cells; or for inhibiting the growth of tissues described above.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for tissue generation activity include, without limitation, those described in: International Patent Publication No. WO95/16035 (bone, cartilage, tendon); International Patent Publication No. WO95/05846 (nerve, neuronal); International Patent Publication No. WO91/07491 (skin, endothelium).

Assays for wound healing activity include, without limitation, those described in: Winter, *Epidermal Wound Healing*, pps. 71–112 (Maibach, H. I and Rovee, D. T., eds.), Year Book Medical Publishers, Inc., Chicago, as modified by Eaglstein and Mertz, J. Invest. Dermatol 71:382–84 (1978).

Activin/Inhibin Activity

A protein of the present invention may also exhibit activin- or inhibin-related activities. Inhibins are characterized by their ability to inhibit the release of follicle stimulating hormone (FSH), while activins and are characterized by their ability to stimulate the release of follicle stimulating hormone (FSH). Thus, a protein of the present invention, alone or in heterodimers with a member of the inhibin α family, may be useful as a contraceptive based on the ability of inhibins to decrease fertility in female mammals and decrease spermatogenesis in male mammals. Administration of sufficient amounts of other inhibins can induce infertility in these mammals. Alternatively, the protein of the invention, as a homodimer or as a heterodimer with other protein subunits of the inhibin-β group, may be useful as a fertility inducing therapeutic, based upon the ability of activin molecules in stimulating FSH release from cells of the anterior pituitary. See, for example, U.S. Pat. No. 4,798,885. A protein of the invention may also be useful for advancement of the onset of fertility in sexually immature mammals, so as to increase the lifetime reproductive performance of domestic animals such as cows, sheep and pigs.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for activin/inhibin activity include, without limitation, those described in: Vale et al., Endocrinology 91:562–572, 1972; Ling et al., Nature 321:779–782, 1986; Vale et al., Nature 321:776–779, 1986; Mason et al., Nature 318:659–663, 1985; Forage et al., Proc. Natl. Acad. Sci. USA 83:3091–3095, 1986.

Chemotactic/Chemokinetic Activity

A protein of the present invention may have chemotactic or chemokinetic activity (e.g., act as a chemokine) for mammalian cells, including, for example, monocytes, fibroblasts, neutrophils, T-cells, mast cells, eosinophils, epithelial and/or endothelial cells. Chemotactic and chemokinetic proteins can be used to mobilize or attract a desired cell population to a desired site of action. Chemotactic or chemokinetic proteins provide particular advantages in treatment of wounds and other trauma to tissues, as well as in treatment of localized infections. For example, attraction of lymphocytes, monocytes or neutrophils to tumors or sites of infection may result in improved immune responses against the tumor or infecting agent.

A protein or peptide has chemotactic activity for a particular cell population if it can stimulate, directly or indirectly, the directed orientation or movement of such cell population. Preferably, the protein or peptide has the ability to directly stimulate directed movement of cells. Whether a particular protein has chemotactic activity for a population of cells can be readily determined by employing such protein or peptide in any known assay for cell chemotaxis.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for chemotactic activity (which will identify proteins that induce or prevent chemotaxis) consist of assays that measure the ability of a protein to induce the migration of cells across a membrane as well as the ability of a protein to induce the adhesion of one cell population to another cell population. Suitable assays for movement and adhesion include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Marguiles, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 6.12, Measurement of alpha and beta Chemokines 6.12.1–6.12.28; Taub et al. J. Clin. Invest. 95:1370–1376, 1995; Lind et al. APMIS 103:140–146, 1995; Muller et al Eur. J. Immunol. 25:1744–1748; Gruber et al. J. of Immunol. 152:5860–5867, 1994; Johnston et al. J. of Immunol. 153:1762–1768, 1994.

Hemostatic and Thrombolytic Activity

A protein of the invention may also exhibit hemostatic or thrombolytic activity. As a result, such a protein is expected to be useful in treatment of various coagulation disorders (including hereditary disorders, such as hemophilias) or to enhance coagulation and other hemostatic events in treating wounds resulting from trauma, surgery or other causes. A protein of the invention may also be useful for dissolving or inhibiting formation of thromboses and for treatment and prevention of conditions resulting therefrom (such as, for example, infarction of cardiac and central nervous system vessels (e.g., stroke).

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assay for hemostatic and thrombolytic activity include, without limitation, those described in: Linet et al., J. Clin. Pharmacol. 26:131–140, 1986; Burdick et al., Thrombosis Res. 45:413–419, 1987; Humphrey et al., Fibrinolysis 5:71–79 (1991); Schaub, Prostaglandins 35:467–474, 1988.

Receptor/Ligand Activity

A protein of the present invention may also demonstrate activity as receptors, receptor ligands or inhibitors or agonists of receptor/ligand interactions. Examples of such receptors and ligands include, without limitation, cytokine receptors and their ligands, receptor kinases and their ligands, receptor phosphatases and their ligands, receptors involved in cell-cell interactions and their ligands (including without limitation, cellular adhesion molecules (such as selectins, integrins and their ligands) and receptor/ligand pairs involved in antigen presentation, antigen recognition and development of cellular and humoral immune responses). Receptors and ligands are also useful for screening of potential peptide or small molecule inhibitors of the relevant receptor/ligand interaction. A protein of the present invention (including, without limitation, fragments of receptors and ligands) may themselves be useful as inhibitors of receptor/ligand interactions.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for receptor-ligand activity include without limitation those described in:Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 7.28, Measurement of Cellular Adhesion under static conditions 7.28.1–7.28.22), Takai et al., Proc. Natl. Acad. Sci. USA

84:6864–6868, 1987; Bierer et al., J. Exp. Med. 168:1145–1156, 1988; Rosenstein et al., J. Exp. Med. 169:149–160 1989; Stoltenborg et al., J. Immunol. Methods 175:59–68, 1994; Stitt et al., Cell 80:661–670, 1995.

Anti-Inflammatory Activity

Proteins of the present invention may also exhibit anti-inflammatory activity. The anti-inflammatory activity may be achieved by providing a stimulus to cells involved in the inflammatory response, by inhibiting or promoting cell-cell interactions (such as, for example, cell adhesion), by inhibiting or promoting chemotaxis of cells involved in the inflammatory process, inhibiting or promoting cell extravasation, or by stimulating or suppressing production of other factors which more directly inhibit or promote an inflammatory response. Proteins exhibiting such activities can be used to treat inflammatory conditions including chronic or acute conditions), including without limitation intimation associated with infection (such as septic shock, sepsis or systemic inflammatory response syndrome (SIRS)), ischemia-reperfusion injury, endotoxin lethality, arthritis, complement-mediated hyperacute rejection, nephritis, cytokine or chemokine-induced lung injury, inflammatory bowel disease, Crohn's disease or resulting from over production of cytokines such as TNF or IL-1. Proteins of the invention may also be useful to treat anaphylaxis and hypersensitivity to an antigenic substance or material.

Tumor Inhibition Activity

In addition to the activities described above for immunological treatment or prevention of tumors, a protein of the invention may exhibit other anti-tumor activities. A protein may inhibit tumor growth directly or indirectly (such as, for example, via ADCC). A protein may exhibit its tumor inhibitory activity by acting on tumor tissue or tumor precursor tissue, by inhibiting formation of tissues necessary to support tumor growth (such as, for example, by inhibiting angiogenesis), by causing production of other factors, agents or cell types which inhibit tumor growth, or by suppressing, eliminating or inhibiting factors, agents or cell types which promote tumor growth.

Other Activities

A protein of the invention may also exhibit one or more of the following additional activities or effects: inhibiting the growth, infection or function of, or killing, infectious agents, including, without limitation, bacteria, viruses, fungi and other parasites; effecting (suppressing or enhancing) bodily characteristics, including, without limitation, height, weight, hair color, eye color, skin, fat to lean ratio or other tissue pigmentation, or organ or body part size or shape (such as, for example, breast augmentation or diminution, change in bone form or shape); effecting biorhythms or caricadic cycles or rhythms; effecting the fertility of male or female subjects; effecting the metabolism, catabolism, anabolism, processing, utilization, storage or elimination of dietary fat, lipid, protein, carbohydrate, vitamins, minerals, cofactors or other nutritional factors or component(s); effecting behavioral characteristics, including, without limitation, appetite, libido, stress, cognition (including cognitive disorders), depression (including depressive disorders) and violent behaviors; providing analgesic effects or other pain reducing effects; promoting differentiation and growth of embryonic stem cells in lineages other than hematopoietic lineages; hormonal or endocrine activity; in the case of enzymes, correcting deficiencies of the enzyme and treating deficiency-related diseases; treatment of hyperproliferative disorders (such as, for example, psoriasis); immunoglobulin-like activity (such as, for example, the ability to bind antigens or complement); and the ability to act as an antigen in a vaccine composition to raise an immune response against such protein or another material or entity which is cross-reactive with such protein.

ADMINISTRATION AND DOSING

A protein of the present invention (from whatever source derived, including without limitation from recombinant and non-recombinant sources) may be used in a pharmaceutical composition when combined with a pharmaceutically acceptable carrier. Such a composition may also contain (in addition to protein and a carrier) diluents, fillers, salts, buffers, stabilizers, solubilizers, and other materials well known in the art. The term "pharmaceutically acceptable" means a non-toxic material that does not interfere with the effectiveness of the biological activity of the active ingredient(s). The characteristics of the carrier will depend on the route of administration. The pharmaceutical composition of the invention may also contain cytokines, lymphokines, or other hematopoietic factors such as M-CSF, GM-CSF, TNF, IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-11, IL-12, IL-13, IL-14, IL-15, IFN, TNF0, TNF1, TNF2, G-CSF, Meg-CSF, thrombopoietin, stem cell factor, and erythropoietin. The pharmaceutical composition may further contain other agents which either enhance the activity of the protein or compliment its activity or use in treatment. Such additional factors and/or agents may be included in the pharmaceutical composition to produce a synergistic effect with protein of the invention, or to minimize side effects. Conversely, protein of the present invention may be included in formulations of the particular cytokine, lymphokine, other hematopoietic factor, thrombolytic or anti-thrombotic factor, or anti-inflammatory agent to minimize side effects of the cytokine, lymphokine, other hematopoietic factor, thrombolytic or anti-thrombotic factor, or anti-inflammatory agent.

A protein of the present invention may be active in multimers (e.g., heterodimers or homodimers) or complexes with itself or other proteins. As a result, pharmaceutical compositions of the invention may comprise a protein of the invention in such multimeric or complexed form.

The pharmaceutical composition of the invention may be in the form of a complex of the protein(s) of present invention along with protein or peptide antigens. The protein and/or peptide antigen will deliver a stimulatory signal to both B and T lymphocytes. B lymphocytes will respond to antigen through their surface immunoglobulin receptor. T lymphocytes will respond to antigen through the T cell receptor (TCR) following presentation of the antigen by MHC proteins. MHC and structurally related proteins including those encoded by class I and class II MHC genes on host cells will serve to present the peptide antigen(s) to T lymphocytes. The antigen components could also be supplied as purified MHC-peptide complexes alone or with co-stimulatory molecules that can directly signal T cells. Alternatively antibodies able to bind surface immunoglobulin and other molecules on B cells as well as antibodies able to bind the TCR and other molecules on T cells can be combined with the pharmaceutical composition of the invention.

The pharmaceutical composition of the invention may be in the form of a liposome in which protein of the present invention is combined, in addition to other pharmaceutically acceptable carriers, with amphipathic agents such as lipids which exist in aggregated form as micelles, insoluble monolayers, liquid crystals, or lamellar layers in aqueous solution. Suitable lipids for liposomal formulation include, without limitation, monoglycerides, diglycerides, sulfatides, lysolecithin, phospholipids, saponin, bile acids, and the like. Preparation of such liposomal formulations is within the level of skill in the art, as disclosed, for example, in U.S. Pat. Nos. 4,235,871; 4,501,728; 4,837,028; and 4,737,323, all of which are incorporated herein by reference.

As used herein, the term "therapeutically effective amount" means the total amount of each active component of the pharmaceutical composition or method that is sufficient to show a meaningful patient benefit, i.e., treatment, healing, prevention or amelioration of the relevant medical condition, or an increase in rate of treatment, healing, prevention or amelioration of such conditions. When applied to an individual active ingredient, administered alone, the term refers to that ingredient alone. When applied to a combination, the term refers to combined amounts of the active ingredients that result in the therapeutic effect, whether administered in combination, serially or simultaneously.

In practicing the method of treatment or use of the present invention, a therapeutically effective amount of protein of the present invention is administered to a mammal having a condition to be treated. Protein of the present invention may be administered in accordance with the method of the invention either alone or in combination with other therapies such as treatments employing cytokines, lymphokines or other hematopoietic factors. When co-administered with one or more cytokines, lymphokines or other hematopoietic factors, protein of the present invention may be administered either simultaneously with the cytokine(s), lymphokine(s), other hematopoietic factor(s), thrombolytic or antithrombotic factors, or sequentially. If administered sequentially, the attending physician will decide on the appropriate sequence of administering protein of the present invention in combination with cytokine(s), lymphokine(s), other hematopoietic factor(s), thrombolytic or antithrombotic factors.

Administration of protein of the present invention used in the pharmaceutical composition or to practice the method of the present invention can be carried out in a variety of conventional ways, such as oral ingestion, inhalation, topical application or cutaneous, subcutaneous, intraperitoneal, parenteral or intravenous injection. Intravenous administration to the patient is preferred.

When a therapeutically effective amount of protein of the present invention is administered orally, protein of the present invention will be in the form of a tablet, capsule, powder, solution or elixir. When administered in tablet form, the pharmaceutical composition of the invention may additionally contain a solid carrier such as a gelatin or an adjuvant. The tablet, capsule, and powder contain from about 5 to 95% protein of the present invention, and preferably from about 25 to 90% protein of the present invention. When administered in liquid form, a liquid carrier such as water, petroleum, oils of animal or plant origin such as peanut oil, mineral oil, soybean oil, or sesame oil, or synthetic oils may be added. The liquid form of the pharmaceutical composition may further contain physiological saline solution, dextrose or other saccharide solution, or glycols such as ethylene glycol, propylene glycol or polyethylene glycol. When administered in liquid form, the pharmaceutical composition contains from about 0.5 to 90% by weight of protein of the present invention, and preferably from about 1 to 50% protein of the present invention.

When a therapeutically effective amount of protein of the present invention is administered by intravenous, cutaneous or subcutaneous injection, protein of the present invention will be in the form of a pyrogen-free, parenterally acceptable aqueous solution. The preparation of such parenterally acceptable protein solutions, having due regard to pH, isotonicity, stability, and the like, is within the skill in the art. A preferred pharmaceutical composition for intravenous, cutaneous, or subcutaneous injection should contain, in addition to protein of the present invention, an isotonic vehicle such as Sodium Chloride Injection, Ringer's Injection, Dextrose Injection, Dextrose and Sodium Chloride Injection, Lactated Ringer's Injection, or other vehicle as known in the art. The pharmaceutical composition of the present invention may also contain stabilizers, preservatives, buffers, antioxidants, or other additives known to those of skill in the art.

The amount of protein of the present invention in the pharmaceutical composition of the present invention will depend upon the nature and severity of the condition being treated, and on the nature of prior treatments which the patient has undergone. Ultimately, the attending physician will decide the amount of protein of the present invention with which to treat each individual patient. Initially, the attending physician will administer low doses of protein of the present invention and observe the patient's response. Larger doses of protein of the present invention may be administered until the optimal therapeutic effect is obtained for the patient, and at that point the dosage is not increased further. It is contemplated that the various pharmaceutical compositions used to practice the method of the present invention should contain about 0.01 μg to about 100 mg (preferably about 0.1 μg to about 10 mg, more preferably about 0.1 μg to about 1 mg) of protein of the present invention per kg body weight.

The duration of intravenous therapy using the pharmaceutical composition of the present invention will vary, depending on the severity of the disease being treated and the condition and potential idiosyncratic response of each individual patient. It is contemplated that the duration of each application of the protein of the present invention will be in the range of 12 to 24 hours of continuous intravenous administration. Ultimately the attending physician will decide on the appropriate duration of intravenous therapy using the pharmaceutical composition of the present invention.

Protein of the invention may also be used to immunize animals to obtain polyclonal and monoclonal antibodies which specifically react with the protein. Such antibodies may be obtained using either the entire protein or fragments thereof as an immunogen. The peptide immunogens additionally may contain a cysteine residue at the carboxyl terminus, and are conjugated to a hapten such as keyhole limpet hemocyanin (KLH). Methods for synthesizing such peptides are known in the art, for example, as in R. P. Merrifield, J. Amer. Chem. Soc. 85, 2149–2154 (1963); J. L. Krstenansky, et al., FEBS Lett. 211, 10 (1987). Monoclonal antibodies binding to the protein of the invention may be useful diagnostic agents for the immunodetection of the protein. Neutralizing monoclonal antibodies binding to the protein may also be useful therapeutics for both conditions associated with the protein and also in the treatment of some forms of cancer where abnormal expression of the protein is involved. In the case of cancerous cells or leukemic cells, neutralizing monoclonal antibodies against the protein may be useful in detecting and preventing the metastatic spread of the cancerous cells, which may be mediated by the protein.

For compositions of the present invention which are useful for bone, cartilage, tendon or ligament regeneration, the therapeutic method includes administering the composition topically, systematically, or locally as an implant or device. When administered, the therapeutic composition for use in this invention is, of course, in a pyrogen-free, physiologically acceptable form. Further, the composition may desirably be encapsulated or injected in a viscous form for delivery to the site of bone, cartilage or tissue damage. Topical administration may be suitable for wound healing and tissue repair. Therapeutically useful agents other than a protein of the invention which may also optionally be included in the composition as described above, may alternatively or additionally, be administered simultaneously or sequentially with the composition in the methods of the invention. Preferably for bone and/or cartilage formation, the composition would include a matrix capable of delivering the protein-containing composition to the site of bone and/or cartilage damage, providing a structure for the developing bone and cartilage and optimally capable of being resorbed into the body. Such matrices may be formed of materials presently in use for other implanted medical applications.

The choice of matrix material is based on biocompatibility, biodegradability, mechanical properties, cosmetic appearance and interface properties. The particular application of the compositions will define the appropriate formulation. Potential matrices for the compositions may be biodegradable and chemically defined calcium sulfate, tricalciumphosphate, hydroxyapatite, polylactic acid, polyglycolic acid and polyanhydrides. Other potential materials are biodegradable and biologically well-defined, such as bone or dermal collagen. Further matrices are comprised of pure proteins or extracellular matrix components. Other potential matrices are nonbiodegradable and chemically defined, such as sintered hydroxyapatite, bioglass, aluminates, or other ceramics. Matrices may be comprised of combinations of any of the above mentioned types of material, such as polylactic acid and hydroxyapatite or collagen and tricalciumphosphate. The bioceramics may be altered in composition, such as in calcium-aluminate-phosphate and processing to alter pore size, particle size, particle shape, and biodegradability.

Presently preferred is a 50:50 (mole weight) copolymer of lactic acid and glycolic acid in the form of porous particles having diameters ranging from 150 to 800 microns. In some applications, it will be useful to utilize a sequestering agent, such as carboxymethyl cellulose or autologous blood clot, to prevent the protein compositions from disassociating from the matrix.

A preferred family of sequestering agents is cellulosic materials such as alkylcelluloses (including hydroxyalkylcelluloses), including methylcellulose, ethylcellulose, hydroxyethylcellulose, hydroxypropylcellulose, hydroxypropyl-methylcellulose, and carboxymethylcellulose, the most preferred being cationic salts of carboxymethylcellulose (CMC). Other preferred sequestering agents include hyaluronic acid, sodium alginate, poly(ethylene glycol), polyoxyethylene oxide, carboxyvinyl polymer and poly(vinyl alcohol). The amount of sequestering agent useful herein is 0.5–20 wt %, preferably 1–10 wt % based on total formulation weight, which represents the amount necessary to prevent desorbtion of the protein from the polymer matrix and to provide appropriate handling of the composition, yet not so much that the progenitor cells are prevented from infiltrating the matrix, thereby providing the protein the opportunity to assist the osteogenic activity of the progenitor cells.

In further compositions, proteins of the invention may be combined with other agents beneficial to the treatment of the bone and/or cartilage defect, wound, or tissue in question. These agents include various growth factors such as epidermal growth factor (EGF), platelet derived growth factor (PDGF), transforming growth factors (TGF-α and TGF-β), and insulin-like growth factor (IGF).

The therapeutic compositions are also presently valuable for veterinary applications. Particularly domestic animals and thoroughbred horses, in addition to humans, are desired patients for such treatment with proteins of the present invention.

The dosage regimen of a protein-containing pharmaceutical composition to be used in tissue regeneration will be determined by the attending physician considering various factors which modify the action of the proteins, e.g., amount of tissue weight desired to be formed, the site of damage, the condition of the damaged tissue, the size of a wound, type of damaged tissue (e.g., bone), the patient's age, sex, and diet, the severity of any infection, time of administration and other clinical factors. The dosage may vary with the type of matrix used in the reconstitution and with inclusion of other proteins in the pharmaceutical composition. For example, the addition of other known growth factors, such as IGF I (insulin like growth factor I), to the final composition, may also effect the dosage. Progress can be monitored by periodic assessment of tissue/bone growth and/or repair, for example, X-rays, histomorphometric determinations and tetracycline labeling.

Polynucleotides of the present invention can also be used for gene therapy. Such polynucleotides can be introduced either in vivo or ex vivo into cells for expression in a mammalian subject. Polynucleotides of the invention may also be administered by other known methods for introduction of nucleic acid into a cell or organism (including, without limitation, in the form of viral vectors or naked DNA).

Cells may also be cultured ex vivo in the presence of proteins of the present invention in order to proliferate or to produce a desired effect on or activity in such cells. Treated cells can then be introduced in vivo for therapeutic purposes.

Patent and literature references cited herein are incorporated by reference as if fully set forth.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 11

(2) INFORMATION FOR SEQ ID NO:1:

-continued

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 432 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
GGTTTGAAAA CTCTGCTTCC TTTGTGAATT TGGTGTTAGG AGTTCTTATT GTTATTCTGC    60
AGCCTTTACT ATTGTCCTTT ATTTACTGAA CACAGTGAAT ACCAAGCACT GTTTATTAGA   120
GGTTAGGAGT AGGGGCAGGT GATTAAAAAA ACAAAAAAGC TAATAATCTC CTCAAGCAAT   180
TTCTGGCCTA ATAGAATTAT AGTAGACAGT GAAGTATCTA AACCCAGGGA ATCAGATTGA   240
GGCACCATGT CCATCGCCTT GAGAATTAAT AGGCTGCATT TCTGGGTTCT CCNTTTTTTT   300
TTTTTTTTTG CCCAACTGAG TCTTTCTGTG GACTTACATG GAACTTCTTA TTCTCTTAAA   360
TCATTAAGTT ACTTGACAAT ATTCTTGGAT TTGGAGAAAC TGGATGTAGG GCCGTATGAA   420
AAAATCATTC GA                                                       432
```

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: protein

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Met Ser Ile Ala Leu Arg Ile Asn Arg Leu His Phe Trp Val Leu Xaa
1               5                   10                  15

Phe Phe Phe Phe Phe Ala Gln Leu Ser Leu Ser Val Asp Leu His Gly
            20                  25                  30

Thr Ser Tyr Ser Leu Lys Ser Leu Ser Tyr Leu Thr Ile Phe Leu Asp
        35                  40                  45

Leu Glu Lys Leu Asp Val Gly Pro Tyr Glu Lys Ile Ile Arg
    50                  55                  60
```

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 219 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

```
ATAGGATACN GTATCTNGCT TTTTTCATTT AAACGTCGNG AGCAATTTTC CCAAGACATA    60
ACAAACTGTC TTNGAAAAAN GGAAAACATT NGGGGCTGTC AGCANAACNG AAAATGTTTT   120
CTGGGTGAGA CACATGTATC TTNGNAATGG GTTGGATTTA GTGTGCTTTA TTTCAATAAA   180
AATTCAGTAT TATAATTTAA AAAAAAAAAA AAAAAAAAA                          219
```

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 501 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
TCCACAGGTG TCCANTCCCA GGTCCAACTG CAGATTTCGA ATTCGGCCTT CATGGCCTAG    60
AGCGACGCGG AGAARAGCTC CGGGTGCCGC GGCACTGCAG CGCTGAGATT CCTTTACAAA   120
GAAACTCAGA GGACCGGGAA GAAAGAATTT CACCTTTGCG ACGTGCTAGA AAATAARGTC   180
GTCTGGGAAA AGGACTGGAG ACACAAGCGC ATCSCAASYY SRGTGAAGGA SAAASNGAKG   240
GANBTAKWWM MGWGSWGAAA AATKTYWWKC AAMMWMGGTA TTTTCCCTTG GATATTAACT   300
TGCATATCTG AAGAAATGGC ATTCCGGACA ATTTGCGTGT TGGTTGGAGT ATTTATTTGT   360
TCTATCTGTG TGAAAGGATC TTCCCAGCCC CAAGCAAGAG TTTATTTAAC ATTTGATGAA   420
CTTCGAGAAA CCAAGACCTC TGAATACTTC AGCCTTTCCC ACCATCCTTT AGACTACAGG   480
ATTTTATTAA TGGATGAAGA T                                             501
```

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: protein

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

```
Met Ala Phe Arg Thr Ile Cys Val Leu Val Gly Val Phe Ile Cys Ser
1               5                   10                  15

Ile Cys Val Lys Gly Ser Ser Gln Pro Gln Ala Arg Val Tyr Leu Thr
            20                  25                  30

Phe Asp Glu Leu Arg Glu Thr Lys Thr Ser Glu Tyr Phe Ser Leu Ser
        35                  40                  45

His His Pro Leu Asp Tyr Arg Ile Leu Leu Met Asp Glu Asp
    50                  55                  60
```

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 302 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

```
CTAGCACTAG ACATGTCATG GTCTTCATGG TGCATATAAA TATATTTAAC TTAACCCAGA    60
TTTTATTTAT ATCTTTATTC ACCTTTTCTT CAAAATCGAT ATGGTGGCTG CAAAACTAGA   120
ATTGTTGCAT CCCTCAATNG AATGAGGGCC ATATCCCTGT GGTATTCCTT TCCTGCTTNG   180
GGGCTTTAGA ATTCTAATTG TCAGTGATTT TGTATATGAA AACAAGTTCC AAATCCACAG   240
CTTTTACGTA GTAAAAGTCA TAAATGCATA TGACAGAATG GCTATCAAAA GAAAAAAAAA   300
AA                                                                  302
```

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 448 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
GGCGAARGCA GCGGCAGGTC GGGAGCAARA TGGCGCTGCG GCCAGGAGCT GGTTCTGGTG      60

GCGGCGGGGC CGCGARGAKY ATR-
RYGYGRK KTYYRYYSKG KKWKSMGGST TCATGTTTCC     120

TGTTGCAGGT GGGATAAGAC CCCCTCAAGG CCTGATGCCG ATGCAGCAAC AAGGATTTCC     180

TATGGTCTCT GTCATGCAGC CTAATATGCA AGGCATTATG GGAATGAATT ACAGCTCTCA     240

GATGTCCCAA GGACCTATTG CTATGCAGGC AGGAATACCA ATGGGACCAA TGCCAGCAGC     300

GGGAATGCCT TACCTAGGAC AAGCACCCTT CCTGGGCATG CGTCCTCCAG GCCCACAGTA     360

CACTCCAGAC ATGCAGAAGC AGTTTGCCGA AGAGCAGCAG AAACGATTTG AACAGCAGCA     420

AAAACTCTTA GAAAAAAAAA AAAAAAAA                                        448
```

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 107 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: protein

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
Met Phe Pro Val Ala Gly Gly Ile Arg Pro Pro Gln Gly Leu Met Pro
1               5                   10                  15

Met Gln Gln Gln Gly Phe Pro Met Val Ser Val Met Gln Pro Asn Met
            20                  25                  30

Gln Gly Ile Met Gly Met Asn Tyr Ser Ser Gln Met Ser Gln Gly Pro
        35                  40                  45

Ile Ala Met Gln Ala Gly Ile Pro Met Gly Pro Met Pro Ala Ala Gly
    50                  55                  60

Met Pro Tyr Leu Gly Gln Ala Pro Phe Leu Gly Met Arg Pro Pro Gly
65                  70                  75                  80

Pro Gln Tyr Thr Pro Asp Met Gln Lys Gln Phe Ala Glu Glu Gln Gln
                85                  90                  95

Lys Arg Phe Glu Gln Gln Gln Lys Leu Leu Glu
            100                 105
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "oligonucleotide"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:9:

```
GNGCCTCAAT CTGATTCCCT GGGTTTAGA                              29
```

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

-continued

```
            ( D ) TOPOLOGY: linear

     ( i i ) MOLECULE TYPE: other nucleic acid
            ( A ) DESCRIPTION: /desc = "oligonucleotide"

     ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GNCCGGAATG CCATTTCTTC AGATATGCA                                    29

( 2 ) INFORMATION FOR SEQ ID NO:11:

     ( i ) SEQUENCE CHARACTERISTICS:
            ( A ) LENGTH: 29 base pairs
            ( B ) TYPE: nucleic acid
            ( C ) STRANDEDNESS: single
            ( D ) TOPOLOGY: linear

     ( i i ) MOLECULE TYPE: other nucleic acid
            ( A ) DESCRIPTION: /desc = "oligonucleotide"

     ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:11:

TNCCATTGGT ATTCCTGCCT GCATAGCAA                                    29
```

What is claimed is:

1. An isolated polynucleotide selected from the group consisting of:

(a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1;

(b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 247 to nucleotide 432;

(c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 328 to nucleotide 432;

(d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146;

(e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146;

(f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146;

(g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146; and

(h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:2.

2. The polynucleotide of claim 1 comprising the nucleotide sequence of SEQ ID NO:1.

3. The polynucleotide of claim 1 comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 247 to nucleotide 432.

4. The polynucleotide of claim 1 comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 328 to nucleotide 432.

5. The polynucleotide of claim 1 comprising the nucleotide sequence of the full length protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146.

6. The polynucleotide of claim 1 encoding the full length protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146.

7. The polynucleotide of claim 1 comprising the nucleotide sequence of the mature protein coding sequence of clone BD372_5 deposited under accession number ATCC 98146.

8. The polynucleotide of claim 1 encoding the mature protein encoded by the cDNA insert of clone BD372_5 deposited under accession number ATCC 98146.

9. The polynucleotide of claim 1 encoding a protein comprising the amino acid sequence of SEQ ID NO:2.

10. A vector comprising a polynucleotide of claim 1 wherein said polynucleotide is operably linked to an expression control sequence.

11. A host cell transformed with a vector of claim 2.

12. The host cell of claim 3, wherein said cell is a mammalian cell.

13. A process for producing a protein, which comprises:

(a) growing a culture of the host cell of claim 3 in a suitable culture medium; and

(b) purifying the protein from the culture.

14. An isolated gene corresponding to the cDNA sequence of SEQ ID NO:1.

* * * * *

HA

US005552281A

United States Patent [19]

Stashenko et al.

[11] **Patent Number: 5,552,281**

[45] **Date of Patent: Sep. 3, 1996**

[54] **HUMAN OSTEOCLAST-SPECIFIC AND -RELATED GENES**

[75] Inventors: **Philip Stashenko**, Norfolk; **Yi-Ping Li**, Boston; **Anne L. Wucherpfennig**, Brookline, all of Mass.

[73] Assignee: **Forsyth Dental Infirmary for Children**, Boston, Mass.

[21] Appl. No.: **392,678**

[22] Filed: **Feb. 23, 1995**

Related U.S. Application Data

[63] Continuation of Ser. No. 45,270, Apr. 6, 1993, abandoned.

[51] **Int. Cl.^{6}** C07H 21/04; C12N 5/10; C12N 15/70; C12Q 1/68

[52] **U.S. Cl.** 435/6; 435/69.1; 435/172.3; 435/252.3; 435/320.1; 536/23.1

[58] **Field of Search** 435/6, 320.1, 252.3, 435/69.1, 172.3; 536/23.1

[56] **References Cited**

PUBLICATIONS

Blair, Harry C., et al., "Extracellular-matrix degradation at acid pH. Avian osteoclast acid collagenase isolation and characterization", *Biochemical Journal* 290(3):873–884 (15 Mar. 1993).

Tezuka, Ken-Ichi, et al., "Identification of osteopontin in isolated rabbit osteoclasts", *Biochemical and Biophysical Research Communications* 186(2):914–916 (31 Jul. 1992).

Tezuka, Ken-Ichi, et al., "Molecular cloning of a possible cysteine proteinase predominantly expressed in osteoclasts", *Journal of Biological Chemistry* 269(2):1106–1108, (14 Jan. 1994).

Horton, Michael A. et al., "Monoclonal Antibodies to Osteoclastomas (Giant Cell Bone Tumors): Definition of Osteoclast-specific Cellular Antigens," *Cancer Research* 45, 5663–5669 (Nov. 1985).

Davies, John et al., "The Osteoclast Functional Antigen, Implicated in the Regulation of Bone Resorption, Is Biochemically Related to the Vitronectin Receptor," *The Journal of Cell Biology* 109, 1817–1826 (Oct. 1989).

Hayman, Alison, R. et al., "Purification and characterization of a tartrate-resistant acid phosphatase from human osteoclastomas," *Biochem. J.* 261, 601–609 (1989).

Sandberg, M. et al., "Localization of the Expression of Types I, III, and IV Collagen, TGF-β1 and c-fos Genes in Developing Human Calvarial Bones," *Developmental Biology* 130, 324–334 (1988).

Sandberg, M. et al., "Enhanced expression of TGF-β and c-fos mRNAs in the growth plates of developing human long bones," *Development* 102, 461–470 (1988).

Ek-Rylander, Barbro et al., "Cloning, Sequence, and Developmental Expression of a Type 5, Tartrate-resistant, Acid Phoshatase of Rat Bone," *The Journal of Biological Chemistry* 266(36), 24684–24689 (Dec. 25, 1991).

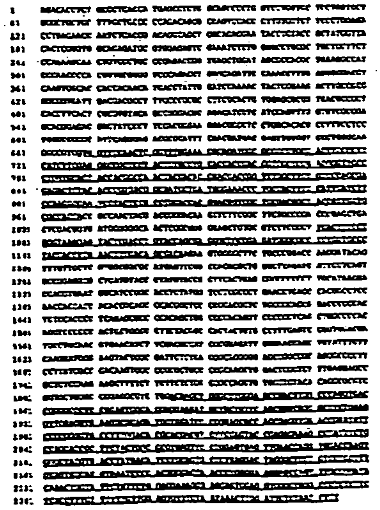
GenBank/EMBL Sequence Search Printout, pp. 1–19 (Jun. 24, 1993).

Primary Examiner—W. Gary Jones
Assistant Examiner—Paul B. Tran
Attorney, Agent, or Firm—Hamilton, Brook, Smith & Reynolds, P.C.

[57] **ABSTRACT**

The present invention relates to purified DNA sequences encoding all or a portion of an osteoclast-specific or -related gene products and a method for identifying such sequences. The invention also relates to antibodies directed against an osteoclast-specific or -related gene product. Also claimed are DNA constructs capable of replicating DNA encoding all or a portion of an osteoclast-specific or -related gene product, and DNA constructs capable of directing expression in a host cell of an osteoclast-specific or -related gene product.

5 Claims, 1 Drawing Sheet



```
1     AGACACCTCT GCCCTCACCA TGAGCCTCTG GCAGCCCCTG GTCCTGGTGC TCCTGGTGCT
61    GGGCTGCTGC TTTGCTGCCC CCAGACAGCG CCAGTCCACC CTTGTGCTCT TCCCTGGAGA
121   CCTGAGAACC AATCTCACCG ACAGGCAGCT GGCAGAGGAA TACCTGTACC GCTATGGTTA
181   CACTCGGGTG GCAGAGATGC GTGGAGAGTC GAAATCTCTG GGGCCTGCGC TGCTGCTTCT
241   CCAGAAGCAA CTGTCCCTGC CCGAGACCGG TGAGCTGGAT AGCGCCACGC TGAAGGCCAT
301   GCGAACCCCA CGGTGCGGGG TCCCAGACCT GGGCAGATTC CAAACCTTTG AGGGCGACCT
361   CAAGTGGCAC CACCACAACA TCACCTATTG GATCCAAAAC TACTCGGAAG ACTTGCCGCG
421   GGCGGTGATT GACGACGCCT TTGCCCGCGC CTTCGCACTG TGGAGCGCGG TGACGCCGCT
481   CACCTTCACT CGCGTGTACA GCCGGGACGC AGACATCGTC ATCCAGTTTG GTGTCGCGGA
541   GCACGGAGAC GGGTATCCCT TCGACGGGAA GGACGGGCTC CTGGCACACG CCTTTCCTCC
601   TGGCCCCGGC ATTCAGGGAG ACGCCCATTT CGACGATGAC GAGTTGTGGT CCCTGGGCAA
661   GGGCGTCGTG GTTCCAACTC GGTTTGGAAA CGCAGATGGC GCGGCCTGCC ACTTCCCCTT
721   CATCTTCGAG GGCCGCTCCT ACTCTGCCTG CACCACCGAC GGTCGCTCCG ACGGGTTGCC
781   CTGGTGCAGT ACCACGGCCA ACTACGACAC CGACGACCGG TTTGGCTTCT GCCCCAGCGA
841   GAGACTCTAC ACCCGGGACG GCAATGCTGA TGGGAAACCC TGCCAGTTTC CATTCATCTT
901   CCAAGGCCAA TCCTACTCCG CCTGCACCAC GGACGGTCGC TCCGACGGCT ACCGCTGGTG
961   CGCCACCACC GCCAACTACG ACCGGGACAA GCTCTTCGGC TTCTGCCCGA CCCGAGCTGA
1021  CTCGACGGTG ATGGGGGGCA ACTCGGCGGG GGAGCTGTGC GTCTTCCCCT TCACTTTCCT
1081  GGGTAAGGAG TACTCGACCT GTACCAGCGA GGGCCGCGGA GATGGGCGCC TCTGGTGCGC
1141  TACCACCTCG AACTTTGACA GCGACAAGAA GTGGGGCTTC TGCCCGGACC AAGGATACAG
1201  TTTGTTCCTC GTGGCGGCGC ATGAGTTCGG CCACGCGCTG GGCTTAGATC ATTCCTCAGT
1261  GCCGGAGGCG CTCATGTACC CTATGTACCG CTTCACTGAG GGGCCCCCCT TGCATAAGGA
1321  CGACGTGAAT GGCATCCGGC ACCTCTATGG TCCTCGCCCT GAACCTGAGC CACGGCCTCC
1381  AACCACCACC ACACCGCAGC CCACGGCTCC CCCGACGGTC TGCCCCACCG GACCCCCCAC
1441  TGTCCACCCC TCAGAGCGCC CCACAGCTGG CCCCACAGGT CCCCCCTCAG CTGGCCCCAC
1501  AGGTCCCCCC ACTGCTGGCC CTTCTACGGC CACTACTGTG CCTTTGAGTC CGGTGGACGA
1561  TGCCTGCAAC GTGAACATCT TCGACGCCAT CGCGGAGATT GGGAACCAGC TGTATTTGTT
1621  CAAGGATGGG AAGTACTGGC GATTCTCTGA GGGCAGGGGG AGCCGGCCGC AGGGCCCCTT
1681  CCTTATCGCC GACAAGTGGC CCGCGCTGCC CCGCAAGCTG GACTCGGTCT TTGAGGAGCC
1741  GCTCTCCAAG AAGCTTTTCT TCTTCTCTGG GCGCCAGGTG TGGGTGTACA CAGGCGCGTC
1801  GGTGCTGGGC CCGAGGCGTC TGGACAAGCT GGGCCTGGGA GCCGACGTGG CCCAGGTGAC
1861  CGGGGCCCTC CGGAGTGGCA GGGGGAAGAT GCTGCTGTTC AGCGGGCGGC GCCTCTGGAG
1921  GTTCGACGTG AAGGCGCAGA TGGTGGATCC CCGGAGCGCC AGCGAGGTGG ACCGGATGTT
1981  CCCCGGGGTG CCTTTGGACA CGCACGACGT CTTCCAGTAC CGAGAGAAAG CCTATTTCTG
2041  CCAGGACCGC TTCTACTGGC GCGTGAGTTC CCGGAGTGAG TTGAACCAGG TGGACCAAGT
2101  GGGCTACGTG ACCTATGACA TCCTGCAGTG CCCTGAGGAC TAGGGCTCCC GTCCTGCTTT
2161  GCAGTGCCAT GTAAATCCCC ACTGGGACCA ACCCTGGGGA AGGAGCCAGT TTGCCGGATA
2221  CAAACTGGTA TTCTGTTCTG GAGGAAAGGG AGGAGTGGAG GTGGGCTGGG CCCTCTCTTC
2281  TCACCTTTGT TTTTTGTTGG AGTGTTTCTA ATAAACTTGG ATTCTCTAAC CTTT
```

Figure 1

HUMAN OSTEOCLAST-SPECIFIC AND -RELATED GENES

RELATED APPLICATION

This application is a continuation of application Ser. No. 08/045,270 filed on Apr. 6, 1993 now abandoned.

BACKGROUND OF THE INVENTION

Excessive bone resorption by osteoclasts contributes to the pathology of many human diseases including arthritis, osteoporosis, periodontitis, and hypercalcemia of malignancy. During resorption, osteoclasts remove both the mineral and organic components of bone (Blair, H. C., et al., *J. Cell Biol.* 102:1164 (1986)). The mineral phase is solubilized by acidification of the sub-osteoclastic lacuna, thus allowing dissolution of hydroxyapatite (Vaes, G., *Clin. Orthop. Relat.* 231:239 (1988)). However, the mechanism(s) by which type I collagen, the major structural protein of bone, is degraded remains controversial. In addition, the regulation of osteoclastic activity is only partly understood. The lack of information concerning osteoclast function is due in part to the fact that these cells are extremely difficult to isolate as pure populations in large numbers. Furthermore, there are no osteoclastic cell lines available. An approach to studying osteoclast function that permits the identification of heretofore unknown osteoclast-specific or -related genes and gene products would allow identification of genes and gene products that are involved in the resorption of bone and in the regulation of osteoclastic activity. Therefore, identification of osteclast-specific or -related genes or gene products would prove useful in developing therapeutic strategies for the treatment of disorders involving aberrant bone resorption.

SUMMARY OF THE INVENTION

The present invention relates to isolated DNA sequences encoding all or a portion of osteoclast-specific or -related gene products. The present invention further relates to DNA constructs capable of replicating DNA encoding osteoclast-specific or -related gene products. In another embodiment, the invention relates to a DNA construct capable of directing expression of all or a portion of the osteoclast-specific or -related gene product in a host cell.

Also encompassed by the present invention are prokaryotic or eukaryotic cells transformed or transfected with a DNA construct encoding all or a portion of an osteoclast-specific or -related gene product. According to a particular embodiment, these cells are capable of replicating the DNA construct comprising the DNA encoding the osteoclast-specific or -related gene product, and, optionally, are capable of expressing the osteoclast-specific or -related gene product. Also claimed are antibodies raised against osteoclast-specific or -related gene products, or portions of these gene products.

The present invention further embraces a method of identifying osteoclast-specific or -related DNA sequences and DNA sequences identified in this manner. In one embodiment, cDNA encoding osteoclast is identified as follows: First, human giant cell tumor of the bone was used to 1) construct a cDNA library; 2) produce ^{32}P-labelled cDNA to use as a stromal cell$^+$; osteoclast$^+$ probe, and 3) produce (by culturing) a stromal cell population lacking osteoclasts. The presence of osteoclasts in the giant cell tumor was confirmed by histological staining for the osteoclast marker, type 5 tartrate-resistant acid phosphatase (TRAP) and with the use of monoclonal antibody reagents.

The stromal cell population lacking osteoclasts was produced by dissociating cells of a giant cell tumor, then growing and passaging the cells in tissue culture until the cell population was homogeneous and appeared fibroblastic. The cultured stromal cell population did not contain osteoclasts. The cultured stromal cells were then used to produce a stromal cell$^+$, osteoclast$^-$ ^{32}P-labelled cDNA probe.

The cDNA library produced from the giant cell tumor of the bone was then screened in duplicate for hybridization to the cDNA probes: one screen was performed with the giant cell tumor cDNA probe (stromal cell$^+$, osteoclast$^+$), while a duplicate screen was performed using the cultured stromal cell cDNA probe (stromal cell$^+$, osteoclast$^-$). Hybridization to a stromal$^+$, osteoclast$^+$ probe, accompanied by failure to hybridize to a stromal$^+$, osteoclast$^-$ probe indicated that a clone contained nucleic acid sequences specifically expressed by osteoclasts.

In another embodiment, genomic DNA encoding osteoclast -specific or -related gene products is identified through known hybridization techniques or amplification techniques. In one embodiment, the present invention relates to a method of identifying DNA encoding an osteoclast-specific or -related protein, or gene product, by screening a cDNA library or a genomic DNA library with a DNA probe comprising one or more sequences selected from the group consisting of the DNA sequences set out in Table I (SEQ ID NOs: 1-32). Finally, the present invention relates to an osteoclast-specific or related protein encoded by a nucleotide sequence comprising a DNA sequence selected from the group consisting of the sequences set out in Table I, or their complementary strands.

BRIEF DESCRIPTION OF FIG. 1

The FIG. 1 shows cDNA sequence (SEQ ID NO: 33) of human gelatinase B, and highlights those portions of the sequence represented by the osteoclast-specific or -related cDNA clones of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

As described herein, Applicant has identified osteoclast-specific or osteoclast-related nucleic acid sequences. These sequences were identified as follows: Human giant cell tumor of the bone was used to 1) construct a cDNA library; 2) produce ^{32}P-labelled cDNA to use as a stromal cell$^+$, osteoclast$^+$probe, and 3) produce (by culturing) a stromal cell population lacking osteoclasts. The presence of osteclasts in the giant cell tumor was confirmed by histological staining for the osteoclast marker, type 5 acid phosphatase (TRAP). In addition, monoclonal antibody reagents were used to characterize the multinucleated cells in the giant cell tumor, which cells were found to have a phenotype distinct from macrophages and consistent with osteoclasts.

The stromal cell population lacking osteoclasts was produced by dissociating cells of a giant cell tumor, then growing the cells in tissue culture for at least five passages. After five passages the cultured cell population was homogeneous and appeared fibroblastic. The cultured population contained no multinucleated cells at this point, tested negative for type 5 acid phosphatase, and tested variably alkaline phosphatase positive. That is, the cultured stromal cell population did not contain osteoclasts. The cultured stromal cells were then used to produce a stromal cell$^+$, osteoclast$^-$ ^{32}P-labelled cDNA probe.

The cDNA library produced from the giant cell tumor of the bone was then screened in duplicate for hybridization to the cDNA probes: one screen was performed with the giant cell tumor cDNA probe (stromal cell$^+$, osteroclast$^+$), while a duplicate screen was performed using the cultured stromal cell cDNA probe (stromal cell$^+$ osteoclast$^-$) Clones that hybridized to the giant cell tumor cDNA probe (stromal$^+$, osteoclast$^+$), but not to the stromal cell cDNA probe (stromal$^+$, osteoclast$^-$), were assumed to contain nucleic acid sequences specifically expressed by osteoclasts.

As a result of the differential screen described herein, DNA specifically expressed in osteoclast cells characterized as described herein was identified. This DNA, and equivalent DNA sequences, is referred to herein as osteoclast-specific or osteoclast-related DNA. Osteoclast-specific or -related DNA of the present invention can be obtained from sources in which it occurs in nature, can be produced recombinantly or synthesized chemically; it can be cDNA, genomic DNA, recombinantly-produced DNA or chemically-produced DNA. An equivalent DNA sequence is one which hybridizes, under standard hybridization conditions, to an osteoclast-specific or -related DNA identified as described herein or to a complement thereof.

Differential screening of a human osteoclastoma cDNA library was performed to identify genes specifically expressed in osteoclasts. Of 12,000 clones screened, 195 clones were identified which are either uniquely expressed in osteoclasts, or are osteoclast-related. These clones were further identified as osteoclast-specific, as evidenced by failure to hybridize to mRNA derived from a variety of unrelated human cell types, including epithelium, fibroblasts, lymphocytes, myelomonocytic cells, osteoblasts, and neuroblastoma cells. Of these, 32 clones contain novel cDNA sequences which were not found in the GenBank database.

A large number of cDNA clones obtained by this procedure were found to represent 92 kDa type IV collagenase (gelatinase B; E.C. 3.4.24.35) as well as tartrate resistant acid phosphatase. In situ hybridization localized mRNA for gelatinase B to multinucleated giant cells in human osteoclastomas. Gelatinase B immunoreactivity was demonstrated in giant cells from 8/8 osteoclastomas, osteoclasts in normal bone, and in osteoclasts of Paget's disease by use of a polyclonal antisera raised against a synthetic gelatinase B peptide. In contrast, no immunoreactivity for 72 kDa type IV collagenase (gelatinase A; E.C. 3.4.24.24), which is the product of a separate gene, was detected in osteoclastomas or normal osteoclasts.

The present invention has utility for the production and identification of nucleic acid probes useful for identifying osteoclast-specific or -related DNA. Osteoclast-specific or -related DNA of the present invention can be used to produce osteoclast-specific or -related gene products useful in the therapeutic treatment of disorders involving aberrant bone resorption. The osteoclast-specific or -related sequences are also useful for generating peptides which can then be used to produce antibodies useful for identifying osteoclast-specific or -related gene products, or for altering the activity of osteoclast-specific or -related gene products. Such antibodies are referred to as osteoclast-specific antibodies. Osteoclast-specific antibodies are also useful for identifying osteoclasts. Finally, osteoclast -specific or -related DNA sequences of the present invention are useful in gene therapy. For example, they can be used to alter the expression in osteoclasts of an aberrant osteoclast -specific or -related gene product or to correct aberrant expression of an osteoclast-specific or -related gene product. The sequences described herein can further be used to cause osteoclast-specific or related gene expression in cells in which such expression does not ordinarily occur, i.e., in cells which are not osteoclasts.

Example 1—Osteoclast cDNA Libary Construction

Messenger RNA (mRNA) obtained from a human osteoclastoma ('giant cell tumor of bone'), was used to construct an osteoclastoma cDNA library. Osteoclastomas are actively bone resorptive tumors, but are usually non-metastatic. In cryostat sections, osteoclastomas consist of ~30% multinucleated cells positive for tartrate resistant acid phosphatase (TRAP), a widely utilized phenotypic marker specific in vivo for osteoclasts (Minkin, *Calcif. Tissue Int.* 34:285–290 (1982)). The remaining cells are uncharacterized 'stromal' cells, a mixture of cell types with fibroblastic/mesenchymal morphology. Although it has not yet been definitively shown, it is generally held that the osteoclasts in these tumors are non-transformed, and are activated to resorb bone in vivo by substance(s) produced by the stromal cell element.

Monoclonal antibody reagents were used to partially characterize the surface phenotype of the multinucleated cells in the giant cell tumors of long bone. In frozen sections, all multinucleated cells expressed CD68, which has previously been reported to define an antigen specific for both osteoclasts and macrophages (Horton, M. A. and M. H. Helfrich, In Biology and Physiology of the Osteoclast, B. R. Rifkin and C. V. Gay, editors, CRC Press, Inc. Boca Raton, Fla., 33–54 (1992)). In contrast, no staining of giant cells was observed for CD11b or CD14 surface antigens, which are present on monocyte/macrophages and granulocytes (Arnaout, M. A. et al. *J. Cell. Physiol.* 137:305 (1988); Haziot, A. et al. *J. Immunol.* 141:547 (1988)). Cytocentrifuge preparations of human peripheral blood monocytes were positive for CD68, CD11b, and CD14. These results demonstrate that the multinucleated giant cells of osteoclastomas have a phenotype which is distinct from that of macrophages, and which is consistent with that of osteoclasts.

Osteoclastoma tissue was snap frozen in liquid nitrogen and used to prepare poly A$^+$ mRA according to standard methods. cDNA cloning into a pcDNAII vector was carried out using a commercially-available kit (Librarian, InVitrogen). Approximately 2.6×10^6 clones were obtained, >95% of which contained inserts of an average length 0.6 kB.

Example 2—Stromal Cell mRNA Preparation

A portion of each osteoclastoma was snap frozen in liquid nitrogen for mRNA preparation. The remainder of the tumor was dissociated using brief trypsinization and mechanical disaggregation, and placed into tissue culture. These cells were expanded in Dulbecco's MEM (high glucose, Sigma) supplemented with 10% newborn calf serum (MA Bioproducts), gentamycin (0.5 mg/ml), 1-glutamine (2 mM) and non-essential amino acids (0.1 mM) (Gibco). The stromal cell population was passaged at least five times, after which it showed a homogenous, fibroblastic looking cell population that contained no multinucleated cells. The stromal cells were mononuclear, tested negative acid phosphatase, and tested variably alkaline phosphatase positive. These findings indicate that propagated stromal cells (i.e., stromal cells that are passaged in culture) are non-osteoclastic and non-activated.

Example 3—Identification of DNA Encoding Osteoclastoma-Specific or -Related Gene Products by Differential screening of an Osteoclastoma cDNA Library

A total of 12,000 clones drawn from the osteoclastoma cDNA library were screened by differential hybridization, using mixed ^{32}P labelled cDNA probes derived from (1) giant cell tumor mRNA (stromal cell$^+$, OC$^+$), and (2) mRNA from stromal cells (stromal cell$^+$, OC$^+$) cultivated from the same tumor. The probes were labelled with ^{32}[P]dCTP by random priming to an activity of ~10^9CPM/μg. Of these 12,000 clones, 195 gave a positive hybridization signal with giant cell (i.e., osteoclast and stromal cell) mRNA, but not with stromal cell mRNA. Additionally, these clones failed to hybridize to cDNA produced from mRNA derived from a variety of unrelated human cell types including epithelial cells, fibroblasts, lymphocytes, myelomonocytic cells, osteoblasts, and neuroblastoma cells. The failure of these clones to hybridize to cDNA produced from mRNA derived from other cell types supports the conclusion that these clones are either uniquely expressed in osteoclasts, or are osteoclast-related.

The osteoclast (OC) cDNA library was screened for differential hybridization to OC cDNA (stromal cell$^+$, OC$^+$) and stromal cell cDNA (stromal cell$^+$, OC$^-$) as follows:

NYTRAN filters (Schleicher & Schuell) were placed on agar plates containing growth medium and ampicillin. Individual bacterial colonies from the OC library were randomly picked and transferred, in triplicate, onto filters with preruled grids and then onto a master agar plate. Up to 200 colonies were inoculated onto a single 90-mm filter/plate using these techniques. The plates were inverted and incubated at 37° C. until the bacterial inoculates had grown (on the filter) to a diameter of 0.5–1.0 mm.

The colonies were then lysed, and the DNA bound to the filters by first placing the filters on top of two pieces of Whatman 3 MM paper saturated with 0.5N NaOH for 5 minutes. The filters were neutralized by placing on two pieces of Whatman 3 MM paper saturated with 1M Tris-HCL, pH 8.0 for 3–5 minutes. Neutralization was followed by incubation on another set of Whatman 3 MM papers saturated with 1M Tris-HCL, pH 8.0/1.5M NaCl for 3–5 minutes. The filters were then washed briefly in 2×SSC.

DNA was immobilized on the filters by baking the filters at 80° C. for 30 minutes. Filters were best used immediately, but they could be stored for up to one week in a vacuum jar at room temperature.

Filters were prehybridized in 5–8 ml of hybridization solution per filter, for 2–4 hours in a heat sealable bag. An additional 2 ml of solution was added for each additional filter added to the hybridization bag. The hybridization buffer consisted of 5×SSC, 5×Denhardt's solution, 1% SDS and 100 μg/ml denatured heterologous DNA.

Prior to hybridization, labeled probe was denatured by heating in 1×SSC for 5 minutes at 100° C., then immediately chilled on ice. Denatured probe was added to the filters in hybridization solution, and the filters hybridized with continuous agitation for 12–20 hours at 65° C.

After hybridization, the filters were washed in 2×SSC/0.2% SDS at 50°–60° C. for 30 minutes, followed by washing in 0.2×SSC/0.2% SDS at 60° C. for 60 minutes. The filters were then air dried and autoradiographed using an intensifying screen at –70° C. overnight.

Example 4—DNA Sequencing of Selected Clones

Clones reactive with the mixed tumor probe, but unreactive with the stromal cell probe, are expected to contain either osteoclast-related, or in vivo 'activated' stromal-cell-related gene products. One hundred and forty-four cDNA clones that hybridized to tumor cell cDNA, but not to stromal cell cDNA, were sequenced by the dideoxy chain termination method of Sanger et al. (Sanger F., et al. *Proc. Natl. Acad. Sci. USA* 74:5463 (1977)) using sequenase (US Biochemical). The DNASIS (Hitatchi) program was used to carry out sequence analysis and a homology search in the GenBank/EMBL database.

Fourteen of the 195 tumor$^+$ stromal$^-$ clones were identified as containing inserts with a sequence identical to the osteoclast marker, type 5 tartrate-resistant acid phosphatase (TRAP) (GenBank accession number J04430 M19534). The high representation of TRAP positive clones also indicates the effectiveness of the screening procedure in enriching for clones which contain osteoclast-specific or related cDNA sequences.

Interestingly, an even larger proportion of the tumor$^+$ stromal$^-$ clones (77/195; 39.5%) were identified as human gelatinase B (macrophage-derived gelatinase) (Wilhelm, S. M. *J. Biol. Chem.* 264:17213 (1989)), again indicating high expression of this enzyme by osteoclasts. Twenty-five of the gelatinase B clones were identified by dideoxy sequence analysis; all 25 showed 100% sequence homology to the published gelatinase B sequence (Genbank accession number J05070). The portions of the gelatinase B cDNA sequence covered by these clones is shown in the FIGURE (SEQ ID NO: 33). An additional 52 gelatinase B clones were identified by reactivity with a ^{32}P-labelled probe for gelatinase B.

Thirteen of the sequenced clones yielded no readable sequence. A DNASIS search of GenBank/EMBL databases revealed that, of the remaining 91 clones, 32 clones contain novel sequences which have not yet been reported in the databases or in the literature. These partial sequences are presented in Table I. Note that three of these sequences were repeats, indicating fairly frequent representation of mRNA related to this sequence. The repeat sequences are indicated by$^{a, b}$ superscripts (Clones 198B, 223B and 32C of Table I).

TABLE I

PARTIAL SEQUENCES OF 32 NOVEL OC-SPECIFIC OR -RELATED EXPRESSED GENES (cDNA CLONES)

34A (SEQ ID NO: 1)						
1	GCAAATATCT	AAGTTTATTG	CTTGGATTTC	TAGTGAGAGC	TGTTGAATTT	GGTGATGTCA
61	AATGTTTCTA	GGGTTTTTTT	AGTTTGTTTT	TATTGAAAAA	TTTAATTATT	TATGCTATAG
121	GTGATATTCT	CTTTGAATAA	ACCTATAATA	GAAAATAGCA	GCAGACAACA	
4B (SEQ ID NO: 2)						
1	GTGTCAACCT	GCATATCCTA	AAAATGTCAA	AATGCTGCAT	CTGGTTAATG	TCGGGGTAGG

TABLE I-continued

PARTIAL SEQUENCES OF 32 NOVEL OC-SPECIFIC OR -RELATED EXPRESSED GENES (cDNA CLONES)

61	GGG					
12B (SEQ ID NO: 3)						
1	CTTCCCTCTC	TTGCTTCCCT	TTCCCAAGCA	GAGGTGCTCA	CTCCATGGCC	ACCGCCACCA
61	CAGGCCCACA	GGGAGTACTG	CCAGACTACT	GCTGATGTTC	TCTTAAGGCC	CAGGGAGTCT
121	CAACCAGCTG	GTGGTGAATG	CTGCCTGGCA	CGGGACCCCC	CCC	
28B (SEQ ID NO: 4)						
1	TTTTATTTGT	AAATATATGT	ATTACATCCC	TAGAAAAAGA	ATCCCAGGAT	TTTCCCTCCT
61	GTGTGTTTTC	GTCTTGCTTC	TTCATGGTCC	ATGATGCCAG	CTGAGGTTGT	CAGTACAATG
121	AAACCAAACT	GGCGGGATGG	AAGCAGATTA	TTCTGCCATT	TTTCCAGGTC	TTT
37B (SEQ ID NO: 5)						
1	GGCTGGACAT	GGGTGCCCTC	CACGTCCCTC	ATATCCCCAG	GCACACTCTG	GCCTCAGGTT
61	TTGCCCTGGC	CATGTCATCT	ACCTGGAGTG	GGCCCTCCCCC	TTCTTCAGCC	TTGAATCAAA
121	AGCCACTTTG	TTAGGCGAGG	ATTTCCCAGA	CCACTCATCA	CATTAAAAAA	TATTTTGAAA
181	ACAAAAAAAA	AAAAAAA				
55B (SEQ ID NO: 6)						
1	TTGACAAAGC	TGTTTATTTC	CACCAATAAA	TAGTATATGG	TGATTGGGGT	TTCTATTTAT
61	AAGAGTAGTG	GCTATTATAT	GGGGTATCAT	GTTGATGCTC	ATAAATAGTT	CATATCTACT
121	TAATTTGCCT	TC				
60B (SEQ ID NO: 7)						
1	GAAGAGAGTT	GTATGTACAA	CCCCAACAGG	CAAGGCAGCT	AAATGCAGAG	GGTACAGAGA
61	GATCCCGAGG	GAATT				
86B (SEQ ID NO: 8)						
1	GGATGGAAAC	ATGTAGAAGT	CCAGAGAAAA	ACAATTTTAA	AAAAAGGTGG	AAAAGTTACG
61	GCAAACCTGA	GATTTCAGCA	TAAAATCTTT	AGTTAGAAGT	GAGAGAAAGA	AGAGGGAGGC
121	TGGTTGCTGT	TGCACGTATC	AATAGGTTAT	C		
87B (SEQ ID NO: 9)						
1	TTCTTGATCT	TTAGAACACT	ATGAATAGGG	AAAAAAGAAA	AAACTGTTCA	AAATAAAATG
61	TAGGAGCCGT	GCTTTTGGAA	TGCTTGAGTG	AGGAGCTCAA	CAAGTCCTCT	CCCAAGAAAG
181	CAATGATAAA	ACTTGACAAA	A			
98B (SEQ ID NO: 10)						
1	ACCCATTTCT	AACAATTTTT	ACTGTAAAAT	TTTTGGTCAA	AGTTCTAAGC	TTAATCACAT
61	CTCAAAGAAT	AGAGGCAATA	TATAGCCCAT	CTTACTAGAC	ATACAGTATT	AAACTGGACT
121	GAATATGAGG	ACAAGCTCTA	GTGGTCATTA	AACCCCTCAG	AA	
110B (SEQ ID NO: 11)						
1	ACATATATTA	ACAGCATTCA	TTTGGCCAAA	ATCTACACGT	TTGTAGAATC	CTACTGTATA
61	TAAAGTGGGA	ATGTATCAAG	TATAGACTAT	GAAAGTGCAA	ATAACAAGTC	AAGGTTAGAT
121	TAACTTTTTT	TTTTTACATT	ATAAAATTAA	CTTGTTT		
118B (SEQ ID NO: 12)						
1	CCAAATTTCT	CTGGAATCCA	TCCTCCCTCC	CATCACCATA	GCCTCGAGAC	GTCATTTCTG
61	TTTGACTACT	CCAGC				
133B (SEQ ID NO: 13)						
1	AACTAACCTC	CTCGGACCCC	TGCCTCACTC	ATTTACACCA	ACCACCCAAC	TATCTATAAA
61	CCTGAGCCAT	GGCCATCCCT	TATGAGCGGC	GCAGTGATTA	TAGGCTTTCG	CTCTAAGATA
121	AAAT					
140B (SEQ ID NO: 14)						
1	ATTATTATTC	TTTTTTTATG	TTAGCTTAGC	CATGCAAAAT	TTACTGGTGA	AGCAGTTAAT
61	AAAACACACA	TCCCATTGAA	GGGTTTTGTA	CATTTCAGTC	CTTACAAATA	ACAAAGCAAT
121	GATAAACCCG	GCACGTCCTG	ATAGGAAATT	C		
144B (SEQ ID NO: 15)						
1	CGTGACACAA	ACATGCATTC	GTTTTATTCA	TAAAACAGCC	TGGTTTCCTA	AAACAATACA
61	AACAGCATGT	TCATCAGCAG	GAAGCTGGCC	GTGGGCAGGG	GGGCC	
198B[a] (SEQ ID NO: 16)						
1	ATAGGTTAGA	TTCTCATTCA	CGGGACTAGT	TAGCTTTAAG	CACCCTAGAG	GACTAGGGTA
61	ATCTGACTTC	TCACTTCCTA	AGTTCCCTCT	TATATCCTCA	AGGTAGAAAT	GTCTATGTTT
121	TCTACTCCAA	TTCATAAATC	TATTCATAAG	TCTTTGGTAC	AAGTTACATG	ATAAAAAGAA
181	ATGTGATTTG	TCTTCCCTTC	TTTGCACTTT	TRAAATAAAG	TATTTATCTC	CTGTCTACAG
241	TTTAAT					
212B (SEQ ID NO: 17)						
1	GTCCAGTATA	AAGGAAAGCG	TTAAGTCGGT	AAGCTAGAGG	ATTGTAAATA	TCTTTTATGT
61	CCTCTAGATA	AAACACCCGA	TTAACAGATG	TTAACCTTTT	ATGTTTTGAT	TTGCTTTAAA
121	AATGGCCTTC	TACACATTAG	CTCCAGCTAA	AAAGACACAT	TGAGAGCTTA	GAGGATAGTC
181	TCTGGAGC					
223B[b] (SEQ ID NO: 18)						
1	GCACTTGGAA	GGGAGTTGGT	GTGCTATTTT	TGAAGCAGAT	GTGGTGATAC	TGAGATTGTC
61	TGTTCAGTTT	CCCCATTTGT	TTGTGCTTCA	AATGATCCTT	CCTACTTTGC	TTCTCTCCAC
121	CCATGACCTT	TTTCACTGTG	GCCATCAAGG	ACTTTCCTGA	CAGCTTGTGT	ACTCTTAGGC
181	TAAGAGATGT	GACTACAGCC	TGCCCCTGAC	TG		
241B (SEQ ID NO: 19)						
1	TGTTAGTTTT	TAGGAAGGCC	TGTCTTCTGG	GAGTGAGGTT	TATTAGTCCA	CTTCTTGGAG
61	CTAGACGTCC	TATAGTTAGT	CACTGGGGAT	GGTGAAAGAG	GGAGAAGAGG	AAGGGCGAAG
121	GGAAGGGCTC	TTTGCTAGTA	TCTCCATTTC	TAGAAGATGG	TTTAGATGAT	AACCACAGGT
181	CTATATGAGC	ATAGTAAGGC	TGT			
32C[a] (SEQ ID NO: 20)						
1	CCTATTTCTG	ATCCTGACTT	TGGACAAGGC	CCTTCAGCCA	GAAGACTGAC	AAAGTCATCC
121	TCCGTCTACC	AGAGCGTGCA	CTTGTGATCC	TAAAATAAGC	TTCATCTCCG	GCTGTGCCTT
161	GGGTGGAAGG	GGCAGGATTC	TGCAGCTGCT	TTTGCATTTC	TCTTCCTAAA	TTTCATT

TABLE I-continued

PARTIAL SEQUENCES OF 32 NOVEL OC-SPECIFIC OR -RELATED EXPRESSED GENES (cDNA CLONES)

34C (SEQ ID NO: 21)						
1	CGGAGCGTAG	GTGTGTTTAT	TCCTGTACAA	ATCATTACAA	AACCAAGTCT	GGGGCAGTCA
61	CCGCCCCCAC	CCATCACCCC	AGTGCAATGG	CTAGCTGCTG	GCCTTT	
47C (SEQ ID NO: 22)						
1	TTAGTTCAGT	CAAAGCAGGC	AACCCCCTTT	GGCACTGCTG	CCACTGGGGT	CATGGCGGTT
61	GTGGCAGCTG	GGGAGGTTTC	CCCAACACCC	TCCTCTGCTT	CCCTGTGTGT	CGGGGTCTCA
121	GGAGCTGACC	CAGAGTGGA				
65C (SEQ ID NO: 23)						
1	GCTGAATGTT	TAAGAGAGAT	TTTGGTCTTA	AAGGCTTCAT	CATGAAAGTG	TACATGCATA
61	TGCAAGTGTG	AATTACGTGG	TATGGATGGT	TGCTTGTTTA	TTAACTAAAG	ATGTACAGCA
121	AACTGCCCGT	TTAGAGTCCT	CTTAATATTG	ATGTCCTAAC	ACTGGGTCTG	CTTATGC
79C (SEQ ID NO: 24)						
1	GGCAGTGGGA	TATGGAATCC	AGAAGGGAAA	CAAGCACTGG	ATAATTAAAA	ACAGCTGGGG
61	AGAAAACTGG	GGAAACAAAG	GATATATCCT	CATGGCTCGA	AATAAGAACA	ACGCCTGTGG
121	CATTGCCAAC	CTGGCCAGCT	TCCCCAAGAT	GTGACTCCAG	CCAGAAA	
84C (SEQ ID NO: 25)						
1	GCCAGGGCGG	ACCGTCTTTA	TTCCTCTCCT	GCCTCAGAGG	TCAGGAAGGA	GGTCTGGCAG
61	GACCTGCAGT	GGGCCCTAGT	CATCTGTGGC	AGCGAAGGTG	AAGGGACTCA	CCTTGTCGCC
121	CGTGCCTGAG	TAGAACTTGT	TCTGGAATTC	C		
86C (SEQ ID NO: 26)						
1	AACTCTTTCA	CACTCTGGTA	TTTTTAGTTT	AACAATATAT	GTGTTGTGTC	TTGGAAATTA
61	GTTCATATCA	ATTCATATTG	AGCTGTCTCA	TTCTTTTTTT	AATGGTCATA	TACAGTAGTA
121	TTCAATTATA	AGAATATATC	CTAATACTTT	TTAAAA		
87C (SEQ ID NO: 27)						
1	GGATAAGAAA	GAAGGCCTGA	GGCCTAGGGG	CCGRGGCTGG	CCTGCGTCTC	AGTCCTGGGA
61	CGCAGCAGCC	CGCACAGGTT	GAGAGGGGCA	CTTCCTCTTG	CTTAGGTTGG	TGAGGATCTG
121	GTCCTGGTTG	GCCGGTGGAG	AGCCACAAAA			
88C (SEQ ID NO: 28)						
1	CTGACCTTCG	AGAGTTTGAC	CTGGAGCCGG	ATACCTACTG	CCGCTATGAC	TCGGTCAGCG
61	TGTTCAACGG	AGCCGTGAGC	GACGACTCCG	GTGGGGAAGT	TCTGCGGCGA	T
89C (SEQ ID NO: 29)						
1	ATCCCTGGCT	GTGGATAGTG	CTTTTGTGTA	GCAAATGCTC	CCTCCTTAAG	GTTATAGGGC
61	TCCCTGAGTT	TGGGAGTGTG	GAAGTACTAC	TTAACTGTCT	GTCCTGCTTG	GCTGTCGTTA
121	TCGTTTTCTG	GTGATGTTGT	GCTAACAATA	AGAATAC		
101C (SEQ ID NO: 30)						
1	GGCTGGGCAT	CCCTCTCCTC	CTCCATCCCC	ATACATCACC	AGGTCTAATG	TTTACAAACG
61	GTGCCAGCCC	GGCTCTGAAG	CCAAGGGCCG	TCCGTGCCAC	GGTGGCTGTG	AGTATTCCTC
121	CGTTAGCTTT	CCCATAAGGT	TGGAGTATCT	GC		
112C (SEQ ID NO: 31)						
1	CCAACTCCTA	CCGCGATACA	GACCCACAGA	GTGCCATCCC	TGAGAGACCA	GACCGCTCCC
161	CAATACTCTC	CTAAAATAAA	CATGAAGCAC			
114C (SEQ ID NO: 32)						
1	CATGGATGAA	TGTCTCATGG	TGGGAAGGAA	CATGGTACAT	TTC	

[a] Repeated 3 times
[b] Repeated 2 times

Sequence analysis of the OC$^+$ stromal cell$^-$ cloned DNA sequences revealed, in addition to the novel sequences, a number of previously-described genes. The known genes identified (including type 5 acid phosphatase, gelatinase B, cystatin C (13 clones), Alu repeat sequences (11 clones), creatine kinase (6 clones) and others) are summarized in Table II. In situ hybridization (described below) directly demonstrated that gelatinase B mRNA is expressed in multinucleated osteoclasts and not in stromal cells. Although gelatinase B is a well-characterized protease, its expression at high levels in osteoclasts has not been previously described. The expression in osteoclasts of cystatin C, a cysteine protease inhibitor, is also unexpected. This finding has not yet been confirmed by in situ hybridization. Taken together, these results demonstrate that most of these identified genes are osteoclast-expressed, thereby confirming the effectiveness of the differential screening strategy for identifying DNA encoding osteoclast-specific or -related gene products. Therefore, novel genes identified by this method have a high probability of being OC-specific or related.

In addition, a minority of the genes identified by this screen are probably not expressed by OCs (Table II). For example, type III collagen (6 clones), collagen type I (1 clone), dermatansulfate (1 clone), and type VI collagen (1 clone) are more likely to originate from the stromal cells or from osteoblastic cells which are present in the tumor. These cDNA sequences survive the differential screening process either because the cells which produce them in the tumor in vivo die out during the stromal cell propagation phase, or because they stop producing their product in vitro. These clones do not constitute more than 5–10% of the all sequences selected by differential hybridization.

TABLE II

SEQUENCE ANALYSIS OF CLONES ENCODING KNOWN SEQUENCES FROM AN OSTEOCLASTOMA cDNA LIBRARY

Clones with Sequence Homology to Collagenase Type IV	25 total
Clones with Sequence Homology to Type 5 Tartrate Resistant Acid Phosphatase	14 total
Clones with Sequence Homology to Cystatin C	13 total
Clones with Sequence Homology to Alu-repeat Sequences	11 total
Clones with Sequence Homology to Creatnine Kinase	6 total
Clones with Sequence Homology to	6 total

TABLE II-continued

SEQUENCE ANALYSIS OF CLONES ENCODING KNOWN SEQUENCES FROM AN OSTEOCLASTOMA cDNA LIBRARY

Type III Collagen	
Clones with Sequence Homology to MHC Class I γ Invariant Chain	5 total
Clones with Sequence Homology to MHC Class II β Chain	3 total
One or Two Clone(s) with Sequence Homology to Each of the Following:	10 total
αl collagen type I	
γ interferon inducible protein	
osteopontin	
Human chondroitin/dermatansulfate	
α globin	
β glucosidase/sphingolipid activator	
Human CAPL protein (Ca binding)	
Human EST 01024	
Type VI collagen	
Human EST 00553	

Example 5—In situ Hybridiation of OC-Expressed Genes

In situ hybridization was performed using probes derived from novel cloned sequences in order to determine whether the novel putative OC-specific or -related genes are differentially expressed in osteoclasts (and not expressed in the stromal cells) of human giant cell tumors. Initially, in situ hybridization was performed using antisense (positive) and sense (negative control) cRNA probes against human type IV collagenase/gelatinase B labelled with ^{35}S-UTP.

A thin section of human giant cell tumor reacted with the antisense probe resulted in intense labelling of all OCs, as indicated by the deposition of silver grains over these cells, but failed to label the stromal cell elements. In contrast, only minimal background labelling was observed with the sense (negative control) probe. This result confirmed that gelatinase B is expressed in human OCs.

In situ hybridization was then carried out using cRNA probes derived from 11/32 novel genes, labelled with digoxigenin UTP according to known methods.

The results of this analysis are summarized in Table III. Clones 28B, 118B, 140B, 198B, and 212B all gave positive reactions with OCs in frozen sections of a giant cell tumor, as did the positive control gelatinase B. These novel clones therefore are expressed in OCs and fulfill all criteria for OC-relatedness. 198B is repeated three times, indicating relatively high expression. Clones 4B, 37B, 88C and 98B produced positive reactions with the tumor tissue; however the signal was not well-localized to OCs. These clones are therefore not likely to be useful and are eliminated from further consideration. Clones 86B and 87B failed to give a positive reaction with any cell type, possibly indicating very low level expression. This group of clones could still be useful but may be difficult to study further. The results of this analysis show that 5/11 novel genes are expressed in OCs, indicating that ~50% of novel sequences likely to be OC-related.

To generate probes for the in situ hybridizations, cDNA derived from novel cloned osteoclast-specific or -related cDNA was subcloned into a BlueScript II SK(−) vector. The orientation of cloned inserts was determined by restriction analysis of subclones. The T7 and T3 promoters in the BlueScriptII vector was used to generate ^{35}S-labelled (^{35}S-UTP 850 Ci/mmol, Amersham, Arlington Heights, Ill.), or UTP digoxygenin labelled cRNA probes.

TABLE III

In Situ HYBRIDIZATION USING PROBES DERIVED FROM NOVEL SEQUENCES

	Reactivity with:	
Clone	Osteoclasts	Stromal Cells
4B	+	+
28B*	+	−
37B	+	+
86B	−	−
87B	−	−
88C	+	+
98B	+	+
118B*	+	−
140B*	+	−
198B*	+	−
212B*	+	−
Gelatinase B*	+	−

*OC-expressed, as indicated by reactivity with antisense probe and lack of reactivity with sense probe on OCs only.

In situ hybridization was carried out on 7 micron cryostat sections of a human osteoclastoma as described previously (Chang, L.-C. et al. *Cancer Res.* 49:6700 (1989)). Briefly, tissue was fixed in 4% paraformaldehyde and embedded in OCT (Miles Inc., Kankakee, Ill.). The sections were rehydrated, postfixed in 4% paraformaldehyde, washed, and pretreated with 10 mM DTT, 10 mM iodoacetamide, 10 mM N-ethylmaleimide and 0.1 triethanolamine-HCL. Prehybridization was done with 50% deionized formamide, 10 mM Tris-HCl, pH 7.0, 1× Denhardt's, 500 mg/ml tRNA, 80 mg/ml salmon sperm DNA, 0.3M NaCl, mM EDTA, and 100 mM DTT at 45° C. for 2 hours. Fresh hybridization solution containing 10% dextran sulfate and 1.5 ng/ml ^{35}S-labelled or digoxygenin labelled RNA probe was applied after heat denaturation. Sections were coverslipped and then incubated in a moistened chamber at 45°–50° C. overnight. Hybridized sections were washed four times with 50% formamide, 2× SSC, containing 10 mM DTT and 0.5% Triton X-100 at 45° C. Sections were treated with RNase A and RNase T1 to digest single-stranded RNA, washed four times in 2× SSC/10 mM DTT.

In order to detect ^{35}S-labelling by autoradiography, slides were dehydrated, dried, and coated with Kodak NTB-2 emulsion. The duplicate slides were split, and each set was placed in a black box with desiccant, sealed, and incubated at 4° C. for 2 days. The slides were developed (4 minutes) and fixed (5 minutes) using Kodak developer D19 and Kodak fixer. Hematoxylin and eosin were used as counterstains.

In order to detect digoxygenin-labelled probes, a Nucleic Acid Detection Kit (Boehringer-Mannheim, Cat. #1175041) was used. Slides were washed in Buffer 1 consisting of 100 mM Tris/150 mM NaCl, pH7.5, for 1 minute. 100 μl Buffer 2 was added (made by adding 2 mg/ml blocking reagent as provided by the manufacturer) in Buffer 1 to each slide. The slides were placed on a shaker and gently swirled at 20° C.

Antibody solutions were diluted 1:100 with Buffer 2 (as provided by the manufacturer). 100 μl of diluted antibody solution was applied to the slides and the slides were then incubated in a chamber for 1 hour at room temperature. The slides were monitored to avoid drying. After incubation with antibody solution, slides were washed in Buffer 1 for 10 minutes, then washed in Buffer 3 containing 2 mM levamisole for 2 minutes.

After washing, 100 μl color solution was added to the slides. Color solution consisted of nitroblue/tetrazolium salt

(NBT) (1:225 dilution) 4.5 μl, 5-bromo-4-chloro-3-indolyl phosphate (1:285 dilution) 3.5 μl, levamisole 0.2 mg in Buffer 3 (as provided by the manufacturer) in a total volume of 1 ml. Color solution was prepared immediately before use.

After adding the color solution, the slides were placed in a dark, humidified chamber at 20° C. for 2-5 hours and monitored for color development. The color reaction was stopped by rinsing slides in TE Buffer.

The slides were stained for 60 seconds in 0.25% methyl green, washed with tap water, then mounted with water-based Permount (Fisher).

Example 6—Immunohistochemistry

Immunohistochemical staining was performed on frozen and paraffin embedded tissues as well as on cytospin preparations (see Table IV). The following antibodies were used: polyclonal rabbit anti-human gelatinase antibodies; Ab110 for gelatinase B; monoclonal mouse anti-human CD68 antibody (clone KP1) (DAKO, Denmark); Mo1 (anti-CD11b) and Mo2 (anti-CD14) derived from ATCC cell lines HB CRL 8026 and TIB 228/HB44. The anti-human gelatinase B antibody Ab110 was raised against a synthetic peptide with the amino acid sequence EALMYPMYRFTEGPPLHK (SEQ ID NO: 34), which is specific for human gelatinase B (Corcoran, M. L. et al. *J. Biol. Chem*, 267:515 (1992)).

Detection of the immunohistochemical staining was achieved by using a goat anti-rabbit glucose oxidase kit (Vector Laboratories, Burlingame Calif.) according to the manufacturer's directions. Briefly, the sections were rehydrated and pretested with either acetone or 0.1% trypsin. Normal goat serum was used to block nonspecific binding. Incubation with the primary antibody for 2 hours or overnight (Ab110:1/500 dilution) was followed by either a glucose oxidase labeled secondary anti-rabbit serum, or, in the case of the mouse monoclonal antibodies, were reacted with purified rabbit anti-mouse Ig before incubation with the secondary antibody.

Paraffin embedded and frozen sections from osteoclastomas (GCT) were reacted with a rabbit antiserum against gelatinase B (antibody 110) (Corcoran, M. L. et al. *J. Biol. chem.* 267:515 (1992)), followed by color development with glucose oxidase linked reagents. The osteoclasts of a giant cell tumor were uniformly strongly positive for gelatinase B, whereas the stromal cells were unreactive. Control sections reacted with rabbit preimmune serum were negative. Identical findings were obtained for all 8 long bone giant cell tumors tested (Table IV). The osteoclasts present in three out of four central giant cell granulomas (GCG) of the mandible were also positive for gelatinase B expression. These neoplasms are similar but not identical to the long bone giant cell tumors, apart from their location in the jaws (Shafer, W. G. et al., Textbook of Oral Pathology, W. B. Saunders Company, Philadelphia, pp. 144–149 (1983)). In contrast, the multinucleated cells from a peripheral giant cell tumor, which is a generally non-resorptive tumor of oral soft tissue, were unreactive with antibody (Shafer, W. G. et. al., Textbook of Oral Pathology, W. B. Saunders Company, Philadelphia, pp. 144–149 (1983)).

Antibody 110 was also utilized to assess the presence of gelatinase B in normal bone (n=3) and in Paget's disease, in which there is elevated bone remodeling and increased osteoclastic activity. Strong staining for gelatinase B was observed in osteoclasts both in normal bone (mandible of a 2 year old), and in Paget's disease. Staining was again absent in controls incubated with preimmune serum. Osteoblasts did not stain in any of the tissue sections, indicating that gelatinase B expression is limited to osteoclasts in bone. Finally, peripheral blood monocytes were also reactive with antibody 110 (Table IV).

TABLE IV

DISTRIBUTION OF GELATINASE B IN VARIOUS TISSUES

Samples	Antibodies tested Ab 110 gelatinase B
GCT frozen (n = 2)	
giant cells	+
stromal cells	–
GCT paraffin (n = 6)	
giant cells	+
stromal cells	–
central GCG (n = 4)	
giant cells	+(¾)
stromal cells	–
peripheral GCT (n – 4)	
giant cells	–
stromal cells	–
Paget's disease (n = 1)	
osteoclasts	+
osteoblasts	–
normal bone (n = 3)	
osteoclasts	+
osteoblasts	–
monocytes (cytospin)	+

Distribution of gelatinase B in multinucleated giant cells, osteoclasts, osteoblasts and stromal cells in various tissues. In general, paraffin embedded tissues were used for these experiments; exceptions are indicated.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments described herein. Such equivalents are intended to be encompassed by the following claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 34

-continued

```
( 2 ) INFORMATION FOR SEQ ID NO:1:

      ( i ) SEQUENCE CHARACTERISTICS:
            ( A ) LENGTH: 170 base pairs
            ( B ) TYPE: nucleic acid
            ( C ) STRANDEDNESS: double
            ( D ) TOPOLOGY: linear

     ( i i ) MOLECULE TYPE: DNA (genomic)

     ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:1:

GCAAATATCT AAGTTTATTG CTTGGATTTC TAGTGAGAGC TGTTGAATTT GGTGATGTCA    60

AATGTTTCTA GGGTTTTTTT AGTTTGTTTT TATTGAAAAA TTTAATTATT TATGCTATAG   120

GTGATATTCT CTTTGAATAA ACCTATAATA GAAAATAGCA GCAGACAACA              170

( 2 ) INFORMATION FOR SEQ ID NO:2:

      ( i ) SEQUENCE CHARACTERISTICS:
            ( A ) LENGTH: 63 base pairs
            ( B ) TYPE: nucleic acid
            ( C ) STRANDEDNESS: double
            ( D ) TOPOLOGY: linear

     ( i i ) MOLECULE TYPE: DNA (genomic)

     ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:2:

GTGTCAACCT GCATATCCTA AAAATGTCAA AATGCTGCAT CTGGTTAATG TCGGGGTAGG    60

GGG                                                                  63

( 2 ) INFORMATION FOR SEQ ID NO:3:

      ( i ) SEQUENCE CHARACTERISTICS:
            ( A ) LENGTH: 163 base pairs
            ( B ) TYPE: nucleic acid
            ( C ) STRANDEDNESS: double
            ( D ) TOPOLOGY: linear

     ( i i ) MOLECULE TYPE: DNA (genomic)

     ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CTTCCCTCTC TTGCTTCCCT TTCCCAAGCA GAGGTGCTCA CTCCATGGCC ACCGCCACCA    60

CAGGCCCACA GGGAGTACTG CCAGACTACT GCTGATGTTC TCTTAAGGCC CAGGGAGTCT   120

CAACCAGCTG GTGGTGAATG CTGCCTGGCA CGGGACCCCC CCC                     163

( 2 ) INFORMATION FOR SEQ ID NO:4:

      ( i ) SEQUENCE CHARACTERISTICS:
            ( A ) LENGTH: 173 base pairs
            ( B ) TYPE: nucleic acid
            ( C ) STRANDEDNESS: double
            ( D ) TOPOLOGY: linear

     ( i i ) MOLECULE TYPE: DNA (genomic)

     ( x i ) SEQUENCE DESCRIPTION: SEQ ID NO:4:

TTTTATTTGT AAATATATGT ATTACATCCC TAGAAAAAGA ATCCCAGGAT TTTCCCTCCT    60

GTGTGTTTTC GTCTTGCTTC TTCATGGTCC ATGATGCCAG CTGAGGTTGT CAGTACAATG   120

AAACCAAACT GGCGGGATGG AAGCAGATTA TTCTGCCATT TTTCCAGGTC TTT          173

( 2 ) INFORMATION FOR SEQ ID NO:5:

      ( i ) SEQUENCE CHARACTERISTICS:
            ( A ) LENGTH: 197 base pairs
            ( B ) TYPE: nucleic acid
            ( C ) STRANDEDNESS: double
```

-continued

(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

```
GGCTGGACAT GGGTGCCCTC CACGTCCCTC ATATCCCCAG GCACACTCTG GCCTCAGGTT    60
TTGCCCTGGC CATGTCATCT ACCTGGAGTG GGCCCTCCCC TTCTTCAGCC TTGAATCAAA   120
AGCCACTTTG TTAGGCGAGG ATTTCCCAGA CCACTCATCA CATTAAAAAA TATTTTGAAA   180
ACAAAAAAAA AAAAAAA                                                  197
```

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 132 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

```
TTGACAAAGC TGTTTATTTC CACCAATAAA TAGTATATGG TGATTGGGGT TTCTATTTAT    60
AAGAGTAGTG GCTATTATAT GGGGTATCAT GTTGATGCTC ATAAATAGTT CATATCTACT   120
TAATTTGCCT TC                                                       132
```

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 75 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
GAAGAGAGTT GTATGTACAA CCCCAACAGG CAAGGCAGCT AAATGCAGAG GGTACAGAGA    60
GATCCCGAGG GAATT                                                     75
```

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 151 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
GGATGGAAAC ATGTAGAAGT CCAGAGAAAA ACAATTTTAA AAAAAGGTGG AAAAGTTACG    60
GCAAACCTGA GATTTCAGCA TAAAATCTTT AGTTAGAAGT GAGAGAAAGA AGAGGGAGGC   120
TGGTTGCTGT TGCACGTATC AATAGGTTAT C                                  151
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 141 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

-continued

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:9:

TTCTTGATCT TTAGAACACT ATGAATAGGG AAAAAAGAAA AAACTGTTCA AAATAAAATG 60

TAGGAGCCGT GCTTTTGGAA TGCTTGAGTG AGGAGCTCAA CAAGTCCTCT CCCAAGAAAG 120

CAATGATAAA ACTTGACAAA A 141

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 162 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:10:

ACCCATTTCT AACAATTTTT ACTGTAAAAT TTTTGGTCAA AGTTCTAAGC TTAATCACAT 60

CTCAAAGAAT AGAGGCAATA TATAGCCCAT CTTACTAGAC ATACAGTATT AAACTGGACT 120

GAATATGAGG ACAAGCTCTA GTGGTCATTA AACCCCTCAG AA 162

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 157 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:11:

ACATATATTA ACAGCATTCA TTTGGCCAAA ATCTACACGT TTGTAGAATC CTACTGTATA 60

TAAAGTGGGA ATGTATCAAG TATAGACTAT GAAAGTGCAA ATAACAAGTC AAGGTTAGAT 120

TAACTTTTTT TTTTTACATT ATAAAATTAA CTTGTTT 157

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 75 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CCAAATTTCT CTGGAATCCA TCCTCCCTCC CATCACCATA GCCTCGAGAC GTCATTTCTG 60

TTTGACTACT CCAGC 75

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 124 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:13:

AACTAACCTC CTCGGACCCC TGCCTCACTC ATTTACACCA ACCACCCAAC TATCTATAAA 60

CCTGAGCCAT GGCCATCCCT TATGAGCGGC GCAGTGATTA TAGGCTTTCG CTCTAAGATA 120

-continued

```
AAAT                                                                     124
```

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 151 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:14:

```
ATTATTATTC TTTTTTTATG TTAGCTTAGC CATGCAAAAT TTACTGGTGA AGCAGTTAAT     60
AAAACACACA TCCCATTGAA GGGTTTTGTA CATTTCAGTC CTTACAAATA ACAAAGCAAT    120
GATAAACCCG GCACGTCCTG ATAGGAAATT C                                   151
```

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 105 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:15:

```
CGTGACACAA ACATGCATTC GTTTTATTCA TAAAACAGCC TGGTTTCCTA AAACAATACA     60
AACAGCATGT TCATCAGCAG GAAGCTGGCC GTGGGCAGGG GGGCC                    105
```

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 246 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:16:

```
ATAGGTTAGA TTCTCATTCA CGGGACTAGT TAGCTTTAAG CACCCTAGAG GACTAGGGTA     60
ATCTGACTTC TCACTTCCTA AGTTCCCTCT TATATCCTCA AGGTAGAAAT GTCTATGTTT    120
TCTACTCCAA TTCATAAATC TATTCATAAG TCTTTGGTAC AAGTTACATG ATAAAAAGAA    180
ATGTGATTTG TCTTCCCTTC TTTGCACTTT TGAAATAAAG TATTTATCTC CTGTCTACAG    240
TTTAAT                                                               246
```

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 188 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:17:

```
GTCCAGTATA AAGGAAAGCG TTAAGTCGGT AAGCTAGAGG ATTGTAAATA TCTTTTATGT     60
CCTCTAGATA AAACACCCGA TTAACAGATG TTAACCTTTT ATGTTTTGAT TTGCTTTAAA    120
AATGGCCTTC TACACATTAG CTCCAGCTAA AAAGACACAT TGAGAGCTTA GAGGATAGTC    180
```

```
TCTGGAGC                                                                  188
```

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 212 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:18:

```
GCACTTGGAA GGGAGTTGGT GTGCTATTTT TGAAGCAGAT GTGGTGATAC TGAGATTGTC      60
TGTTCAGTTT CCCCATTTGT TTGTGCTTCA AATGATCCTT CCTACTTTGC TTCTCTCCAC     120
CCATGACCTT TTTCACTGTG GCCATCAAGG ACTTTCCTGA CAGCTTGTGT ACTCTTAGGC     180
TAAGAGATGT GACTACAGCC TGCCCCTGAC TG                                   212
```

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 203 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:19:

```
TGTTAGTTTT TAGGAAGGCC TGTCTTCTGG GAGTGAGGTT TATTAGTCCA CTTCTTGGAG      60
CTAGACGTCC TATAGTTAGT CACTGGGGAT GGTGAAAGAG GGAGAAGAGG AAGGGCGAAG     120
GGAAGGGCTC TTTGCTAGTA TCTCCATTTC TAGAAGATGG TTTAGATGAT AACCACAGGT     180
CTATATGAGC ATAGTAAGGC TGT                                             203
```

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 177 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:20:

```
CCTATTTCTG ATCCTGACTT TGGACAAGGC CCTTCAGCCA GAAGACTGAC AAAGTCATCC      60
TCCGTCTACC AGAGCGTGCA CTTGTGATCC TAAAATAAGC TTCATCTCCG GCTGTGCCTT     120
GGGTGGAAGG GGCAGGATTC TGCAGCTGCT TTTGCATTTC TCTTCCTAAA TTTCATT        177
```

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 106 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:21:

```
CGGAGCGTAG GTGTGTTTAT TCCTGTACAA ATCATTACAA AACCAAGTCT GGGGCAGTCA      60
CCGCCCCCAC CCATCACCCC AGTGCAATGG CTAGCTGCTG GCCTTT                    106
```

-continued

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 139 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:22:

```
TTAGTTCAGT CAAAGCAGGC AACCCCCTTT GGCACTGCTG CCACTGGGGT CATGGCGGTT    60
GTGGCAGCTG GGGAGGTTTC CCCAACACCC TCCTCTGCTT CCCTGTGTGT CGGGGTCTCA   120
GGAGCTGACC CAGAGTGGA                                                139
```

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 177 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:23:

```
GCTGAATGTT TAAGAGAGAT TTTGGTCTTA AAGGCTTCAT CATGAAAGTG TACATGCATA    60
TGCAAGTGTG AATTACGTGG TATGGATGGT TGCTTGTTTA TTAACTAAAG ATGTACAGCA   120
AACTGCCCGT TTAGAGTCCT CTTAATATTG ATGTCCTAAC ACTGGGTCTG CTTATGC      177
```

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 167 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:24:

```
GGCAGTGGGA TATGGAATCC AGAAGGGAAA CAAGCACTGG ATAATTAAAA ACAGCTGGGG    60
AGAAAACTGG GGAAACAAAG GATATATCCT CATGGCTCGA AATAAGAACA ACGCCTGTGG   120
CATTGCCAAC CTGGCCAGCT TCCCCAAGAT GTGACTCCAG CCAGAAA                 167
```

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 151 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:25:

```
GCCAGGGCGG ACCGTCTTTA TTCCTCTCCT GCCTCAGAGG TCAGGAAGGA GGTCTGGCAG    60
GACCTGCAGT GGGCCCTAGT CATCTGTGGC AGCGAAGGTG AAGGGACTCA CCTTGTCGCC   120
CGTGCCTGAG TAGAACTTGT TCTGGAATTC C                                  151
```

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 156 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:26:

```
AACTCTTTCA CACTCTGGTA TTTTTAGTTT AACAATATAT GTGTTGTGTC TTGGAAATTA    60
GTTCATATCA ATTCATATTG AGCTGTCTCA TTCTTTTTTT AATGGTCATA TACAGTAGTA   120
TTCAATTATA AGAATATATC CTAATACTTT TTAAAA                             156
```

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 150 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:27:

```
GGATAAGAAA GAAGGCCTGA GGGCTAGGGG CCGGGGCTGG CCTGCGTCTC AGTCCTGGGA    60
CGCAGCAGCC CGCACAGGTT GAGAGGGGCA CTTCCTCTTG CTTAGGTTGG TGAGGATCTG   120
GTCCTGGTTG GCCGGTGGAG AGCCACAAAA                                    150
```

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 212 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:28:

```
GCACTTGGAA GGGAGTTGGT GTGCTATTTT TGAAGCAGAT GTGGTGATAC TGAGATTGTC    60
TGTTCAGTTT CCCCATTTGT TTGTGCTTCA AATGATCCTT CCTACTTTGC TTCTCTCCAC   120
CCATGACCTT TTTCACTGTG GCCATCAAGG ACTTTCCTGA CAGCTTGTGT ACTCTTAGGC   180
TAAGAGATGT GACTACAGCC TGCCCCTGAC TG                                 212
```

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 157 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA (genomic)

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:29:

```
ATCCCTGGCT GTGGATAGTG CTTTTGTGTA GCAAATGCTC CCTCCTTAAG GTTATAGGGC    60
TCCCTGAGTT TGGGAGTGTG GAAGTACTAC TTAACTGTCT GTCCTGCTTG GCTGTCGTTA   120
TCGTTTTCTG GTGATGTTGT GCTAACAATA AGAATAC                            157
```

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 152 base pairs
(B) TYPE: nucleic acid

-continued

(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

GGCTGGGCAT CCCTCTCCTC CTCCATCCCC ATACATCACC AGGTCTAATG TTTACAAACG 60

GTGCCAGCCC GGCTCTGAAG CCAAGGGCCG TCCGTGCCAC GGTGGCTGTG AGTATTCCTC 120

CGTTAGCTTT CCCATAAGGT TGGAGTATCT GC 152

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 90 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

CCAACTCCTA CCGCGATACA GACCCACAGA GTGCCATCCC TGAGAGACCA GACCGCTCCC 60

CAATACTCTC CTAAAATAAA CATGAAGCAC 90

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 43 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

CATGGATGAA TGTCTCATGG TGGGAAGGAA CATGGTACAT TTC 43

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2333 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

AGACACCTCT GCCCTCACCA TGAGCCTCTG GCAGCCCCTG GTCCTGGTGC TCCTGGTGCT 60

GGGCTGCTGC TTTGCTGCCC CCAGACAGCG CCAGTCCACC CTTGTGCTCT TCCCTGGAGA 120

CCTGAGAACC AATCTCACCG ACAGGCAGCT GGCAGAGGAA TACCTGTACC GCTATGGTTA 180

CACTCGGGTG GCAGAGATGC GTGGAGAGTC GAAATCTCTG GGGCCTGCGC TGCTGCTTCT 240

CCAGAAGCAA CTGTCCCTGC CCGAGACCGG TGAGCTGGAT AGCGCCACGC TGAAGGCCAT 300

GCGAACCCCA CGGTGCGGGG TCCCAGACCT GGGCAGATTC CAAACCTTTG AGGGCGACCT 360

CAAGTGGCAC CACCACAACA TCACCTATTG GATCCAAAAC TACTCGGAAG ACTTGCCGCG 420

GGCGGTGATT GACGACGCCT TTGCCCGCGC CTTCGCACTG TGGAGCGCGG TGACGCCGCT 480

CACCTTCACT CGCGTGTACA GCCGGGACGC AGACATCGTC ATCCAGTTTG GTGTCGCGGA 540

GCACGGAGAC GGGTATCCCT TCGACGGGAA GGACGGGCTC CTGGCACACG CCTTTCCTCC 600

TGGCCCCGGC ATTCAGGGAG ACGCCCATTT CGACGATGAC GAGTTGTGGT CCCTGGGCAA 660

-continued

```
GGGCGTCGTG GTTCCAACTC GGTTTGGAAA CGCAGATGGC GCGGCCTGCC ACTTCCCCTT    720
CATCTTCGAG GGCCGCTCCT ACTCTGCCTG CACCACCGAC GGTCGCTCCG ACGGGTTGCC    780
CTGGTGCAGT ACCACGGCCA ACTACGACAC CGACGACCGG TTTGGCTTCT GCCCCAGCGA    840
GAGACTCTAC ACCCGGGACG GCAATGCTGA TGGGAAACCC TGCCAGTTTC CATTCATCTT    900
CCAAGGCCAA TCCTACTCCG CCTGCACCAC GGACGGTCGC TCCGACGGCT ACCGCTGGTG    960
CGCCACCACC GCCAACTACG ACCGGGACAA GCTCTTCGGC TTCTGCCCGA CCCGAGCTGA   1020
CTCGACGGTG ATGGGGGGCA ACTCGGCGGG GGAGCTGTGC GTCTTCCCCT TCACTTTCCT   1080
GGGTAAGGAG TACTCGACCT GTACCAGCGA GGGCCGCGGA GATGGGCGCC TCTGGTGCGC   1140
TACCACCTCG AACTTTGACA GCGACAAGAA GTGGGGCTTC TGCCCGGACC AAGGATACAG   1200
TTTGTTCCTC GTGGCGGCGC ATGAGTTCGG CCACGCGCTG GGCTTAGATC ATTCCTCAGT   1260
GCCGGAGGCG CTCATGTACC CTATGTACCG CTTCACTGAG GGGCCCCCCT TGCATAAGGA   1320
CGACGTGAAT GGCATCCGGC ACCTCTATGG TCCTCGCCCT GAACCTGAGC CACGGCCTCC   1380
AACCACCACC ACACCGCAGC CCACGGCTCC CCCGACGGTC TGCCCCACCG GACCCCCCAC   1440
TGTCCACCCC TCAGAGCGCC CCACAGCTGG CCCCACAGGT CCCCCCTCAG CTGGCCCCAC   1500
AGGTCCCCCC ACTGCTGGCC CTTCTACGGC CACTACTGTG CCTTTGAGTC CGGTGGACGA   1560
TGCCTGCAAC GTGAACATCT TCGACGCCAT CGCGGAGATT GGGAACCAGC TGTATTTGTT   1620
CAAGGATGGG AAGTACTGGC GATTCTCTGA GGGCAGGGGG AGCCGGCCGC AGGGCCCCTT   1680
CCTTATCGCC GACAAGTGGC CCGCGCTGCC CCGCAAGCTG GACTCGGTCT TTGAGGAGCC   1740
GCTCTCCAAG AAGCTTTTCT TCTTCTCTGG GCGCCAGGTG TGGGTGTACA CAGGCGCGTC   1800
GGTGCTGGGC CCGAGGCGTC TGGACAAGCT GGGCCTGGGA GCCGACGTGG CCCAGGTGAC   1860
CGGGGCCCTC CGGAGTGGCA GGGGGAAGAT GCTGCTGTTC AGCGGGCGGC GCCTCTGGAG   1920
GTTCGACGTG AAGGCGCAGA TGGTGGATCC CCGGAGCGCC AGCGAGGTGG ACCGGATGTT   1980
CCCCGGGGTG CCTTTGGACA CGCACGACGT CTTCCAGTAC CGAGAGAAAG CCTATTTCTG   2040
CCAGGACCGC TTCTACTGGC GCGTGAGTTC CCGGAGTGAG TTGAACCAGG TGGACCAAGT   2100
GGGCTACGTG ACCTATGACA TCCTGCAGTG CCCTGAGGAC TAGGGCTCCC GTCCTGCTTT   2160
GCAGTGCCAT GTAAATCCCC ACTGGGACCA ACCCTGGGGA AGGAGCCAGT TTGCCGGATA   2220
CAAACTGGTA TTCTGTTCTG GAGGAAAGGG AGGAGTGGAG GTGGGCTGGG CCCTCTCTTC   2280
TCACCTTTGT TTTTTGTTGG AGTGTTTCTA ATAAACTTGG ATTCTCTAAC CTTT         2334
```

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 18 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: unknown

(i i) MOLECULE TYPE: peptide

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:34:

```
Glu Ala Leu Met Tyr Pro Met Tyr Arg Phe Thr Gln Gly Pro Pro Leu
1               5                   10                  15

His Lys
```

We claim:

1. An isolated osteoclast-specific or -related DNA sequence, or its complementary sequence, the DNA sequence comprising a nucleic acid sequence selected from the group consisting of:

a) DNA sequences set forth in the group consisting of SEQ ID NOS. 12, 14, 16 and 17, or their complementary strands; and

b) DNA sequences which hybridize under standard conditions to the DNA sequences defined in a).

2. A DNA construct capable of replicating, in a host cell, osteoclast-specific or -related DNA, said construct comprising:

a) a DNA sequence of claim 1; and

b) sequences, in addition to said DNA sequence, necessary for transforming or transfecting a host cell, and for replicating, in a host cell, said DNA sequence.

3. A DNA construct capable or replicating and expressing, in a host cell, osteoclast-specific or -related DNA, said construct comprising:

a) a DNA sequence of claim 2; and

b) sequences, in addition to said DNA sequence, necessary for transforming or transfecting a host cell, and for replicating and expressing, in a host cell, said DNA sequence.

4. A cell stably transformed or transfected with a DNA construct according to claim 3.

5. A cell stably transformed or transfected with a DNA construct according to claim 4.

* * * * *

II

US006340583B1

(12) **United States Patent**
Yan et al.

(10) **Patent No.:** US 6,340,583 B1
(45) **Date of Patent:** Jan. 22, 2002

(54) **ISOLATED HUMAN KINASE PROTEINS, NUCLEIC ACID MOLECULES ENCODING HUMAN KINASE PROTEINS, AND USES THEREOF**

(75) Inventors: Chunhua Yan, Boyds; Karen A. Ketchum, Germantown; Valentina Di Francesco, Rockville; Ellen M. Beasley, Darnestown, all of MD (US)

(73) Assignee: PE Corporation (NY), Norwalk, CT (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: 09/813,817

(22) Filed: Mar. 22, 2001

(51) Int. Cl.^{7} C12N 9/12; C12N 1/20; C12N 15/00; C12N 5/00; C07H 21/04

(52) U.S. Cl. 435/194; 435/320.1; 435/252.3; 435/325; 536/23.2

(58) Field of Search 435/194, 252.3, 435/325, 320.1; 536/23.2

(56) References Cited

PUBLICATIONS

GenEmbl Database, Accession No. D45906, Feb. 1999.*

Sambrook et al., Molecular Cloning Manual, 2nd edition, Cold Spring Harbor Laboratory Press, 1989.*

* cited by examiner

Primary Examiner—Rebecca E. Prouty
Assistant Examiner—M. Monshipouri
(74) *Attorney, Agent, or Firm*—Celera Genomics; Robert A. Millman; Justin D. Karjala

(57) **ABSTRACT**

The present invention provides amino acid sequences of peptides that are encoded by genes within the human genome, the kinase peptides of the present invention. The present invention specifically provides isolated peptide and nucleic acid molecules, methods of identifying orthologs and paralogs of the kinase peptides, and methods of identifying modulators of the kinase peptides.

9 Claims, 41 Drawing Sheets

```
   1 CCCAGGGCGC CGTAGGCGGT GCATCCCGTT CGCGCCTGGG GCTGTGGTCT
  51 TCCCGCGCCT GAGGCGGCGG CGGCAGGAGC TGAGGGGAGT TGTAGGGAAC
 101 TGAGGGGAGC TGCTGTGTCC CCCGCCTCCT CCTCCCCATT TCCGCGCTCC
 151 CGGGACCATG TCCGCGCTGG CGGGTGAAGA TGTCTGGAGG TGTCCAGGCT
 201 GTGGGGACCA CATTGCTCCA AGCCAGATAT GGTACAGGAC TGTCAACGAA
 251 ACCTGGCACG GCTCTTGCTT CCGGTGAAAG TGATGCGCAG CCTGGACCAC
 301 CCCAATGTGC TCAAGTTCAT TGGTGTGCTG TACAAGGATA AGAAGCTGAA
 351 CCTGCTGACA GAGTACATTG AGGGGGGCAC ACTGAAGGAC TTTCTGCGCA
 401 GTATGGATCC GTTCCCCTGG CAGCAGAAGG TCAGGTTTGC CAAAGGAATC
 451 GCCTCCGGAA TGGACAAGAC TGTGGTGGTG GCAGACTTTG GGCTGTCACG
 501 GCTCATAGTG GAAGAGAGGA AAAGGGCCCC CATGGAGAAG GCCACCACCA
 551 AGAAACGCAC CTTGCGCAAG AACGACCGCA AGAAGCGCTA CACGGTGGTG
 601 GGAAACCCCT ACTGGATGGC CCCTGAGATG CTGAACGGAA AGAGCTATGA
 651 TGAGACGGTG GATATCTTCT CCTTTGGGAT CGTTCTCTGT GAGATCATTG
 701 GGCAGGTGTA TGCAGATCCT GACTGCCTTC CCCGAACACT GGACTTTGGC
 751 CTCAACGTGA AGCTTTTCTG GGAGAAGTTT GTTCCCACAG ATTGTCCCCC
 801 GGCCTTCTTC CCGCTGGCCG CCATCTGCTG CAGACTGGAG CCTGAGAGCA
 851 GACCAGCATT CTCGAAATTG GAGGACTCCT TTGAGGCCCT CTCCCTGTAC
 901 CTGGGGGAGC TGGGCATCCC GCTGCCTGCA GAGCTGGAGG AGTTGGACCA
 951 CACTGTGAGC ATGCAGTACG GCCTGACCCG GGACTCACCT CCCTAGCCCT
1001 GGCCCAGCCC CCTGCAGGGG GGTGTTCTAC AGCCAGCATT GCCCCTCTGT
1051 GCCCCATTCC TGCTGTGAGC AGGGCCGTCC GGGCTTCCTG TGGATTGGCG
1101 GAATGTTTAG AAGCAGAACA AACCATTCCT ATTACCTCCC CAGGAGGCAA
1151 GTGGGCGCAG CACCAGGGAA ATGTATCTCC ACAGGTTCTG GGGCCTAGTT
1201 ACTGTCTGTA AATCCAATAC TTGCCTGAAA GCTGTGAAGA AGAAAAAAAC
1251 CCCTGGCCTT TGGGCCAGGA GGAATCTGTT ACTCGAATCC ACCCAGGAAC
1301 TCCCTGGCAG TGGATTGTGG GAGGCTCTTG CTTACACTAA TCAGCGTGAC
1351 CTGGACCTGC TGGGCAGGAT CCCAGGGTGA ACCTGCCTGT GAACTCTGAA
1401 GTCACTAGTC CAGCTGGGTG CAGGAGGACT TCAAGTGTGT GGACGAAAGA
1451 AAGACTGATG GCTCAAAGGG TGTGAAAAAG TCAGTGATGC TCCCCCTTTC
1501 TACTCCAGAT CCTGTCCTTC CTGGAGCAAG GTTGAGGGAG TAGGTTTTGA
1551 AGAGTCCCTT AATATGTGGT GGAACAGGCC AGGAGTTAGA GAAAGGGCTG
1601 GCTTCTGTTT ACCTGCTCAC TGGCTCTAGC CAGCCCAGGG ACCACATCAA
1651 TGTGAGAGGA AGCCTCCACC TCATGTTTTC AAACTTAATA CTGGAGACTG
1701 GCTGAGAACT TACGGACAAC ATCCTTTCTG TCTGAAACAA ACAGTCACAA
1751 GCACAGGAAG AGGCTGGGGG ACTAGAAAGA GGCCCTGCCC TCTAGAAAGC
1801 TCAGATCTTG GCTTCTGTTA CTCATACTCG GGTGGGCTCC TTAGTCAGAT
1851 GCCTAAAACA TTTTGCCTAA AGCTCGATGG GTTCTGGAGG ACAGTGTGGC
1901 TTGTCACAGG CCTAGAGTCT GAGGGAGGGG AGTGGGAGTC TCAGCAATCT
1951 CTTGGTCTTG GCTTCATGGC AACCACTGCT CACCCTTCAA CATGCCTGGT
2001 TTAGGCAGCA GCTTGGGCTG GGAAGAGGTG GTGGCAGAGT CTCAAAGCTG
2051 AGATGCTGAG AGAGATAGCT CCCTGAGCTG GGCCATCTGA CTTCTACCTC
2101 CCATGTTTGC TCTCCCAACT CATTAGCTCC TGGGCAGCAT CCTCCTGAGC
2151 CACATGTGCA GGTACTGGAA AACCTCCATC TTGGCTCCCA GAGCTCTAGG
2201 AACTCTTCAT CACAACTAGA TTTGCCTCTT CTAAGTGTCT ATGAGCTTGC
2251 ACCATATTTA ATAAATTGGG AATGGGTTTG GGGTATTAAA AAAAAAAAAA
2301 AAAAAAAAAA AAAAAAAAAA  (SEQ ID NO:1)
```

FIG.1A

```
FEATURES:
5'UTR:          1-228
Start Codon:    229
Stop Codon:     994
3'UTR:          997
```

Homologous proteins:
Top 10 BLAST Hits

```
                                                                             Score   E
CRA|1000682328847 /altid=gi|8051618 /def=ref|NP_057952.1| LIM d...            485  e-136
CRA|18000005015874 /altid=gi|5031869 /def=ref|NP_005560.1| LIM ...            485  e-136
CRA|88000001156379 /altid=gi|7434382 /def=pir||JC5814 LIM motif...            469  e-131
CRA|88000001156378 /altid=gi|7434381 /def=pir||JC5813 LIM motif...            469  e-131
CRA|18000005154371 /altid=gi|7428032 /def=pir||JE0240 LIM kinas...            469  e-131
CRA|18000005126937 /altid=gi|6754550 /def=ref|NP_034848.1| LIM ...            469  e-131
CRA|18000005127186 /altid=gi|2804562 /def=dbj|BAA24491.1| (AB00...            469  e-131
CRA|18000005127185 /altid=gi|2804553 /def=dbj|BAA24489.1| (AB00...            469  e-131
CRA|18000005004416 /altid=gi|2143830 /def=pir||I78847 LIM motif...            468  e-131
CRA|18000005004415 /altid=gi|1708825 /def=sp|P53670|LIK2_RAT LI...            468  e-131
```

BLAST dbEST hits:

```
                                                                             Score   E
gi|10950740 /dataset=dbest /taxon=96...                                       1049  0.0
gi|10156485 /dataset=dbest /taxon=96...                                        975  0.0
gi|5421647 /dataset=dbest /taxon=9606 ...                                      952  0.0
gi|10895718 /dataset=dbest /taxon=96...                                        757  0.0
gi|13043102 /dataset=dbest /taxon=960...                                       714  0.0
gi|519615 /dataset=dbest /taxon=9606 /...                                      531  e-149
gi|11002869 /dataset=dbest /taxon=96...                                        511  e-143
```

EXPRESSION INFORMATION FOR MODULATORY USE:
library source:
From BLAST dbEST hits:

```
gi|10950740  teratocarcinoma
gi|10156485  ovary
gi|5421647   testis
gi|10895718  nervous_normal
gi|13043102  bladder
gi|519615    infant brain
gi|11002869  thyroid gland
```

From tissue screening panels:
Fetal whole brain

FIG.1B

```
  1 MVQDCQRNLA RLLLPVKVMR SLDHPNVLKF IGVLYKDKKL NLLTEYIEGG
 51 TLKDFLRSMD PFPWQQKVRF AKGIASGMDK TVVVADFGLS RLIVEERKRA
101 PMEKATTKKR TLRKNDRKKR YTVVGNPYWM APEMLNGKSY DETVDIFSFG
151 IVLCEIIGQV YADPDCLPRT LDFGLNVKLF WEKFVPTDCP PAFFPLAAIC
201 CRLEPESRPA FSKLEDSFEA LSLYLGELGI PLPAELEELD HTVSMQYGLT
251 RDSPP  (SEQ ID NO:2)
```

```
FEATURES:
Functional domains and key regions:
[1] PDOC00004 PS00004 CAMP_PHOSPHO_SITE
cAMP- and cGMP-dependent protein kinase phosphorylation site

Number of matches: 2
     1     108-111 KKRT

     2     119-122 KRYT

 [2] PDOC00005 PS00005 PKC_PHOSPHO_SITE
Protein kinase C phosphorylation site

Number of matches: 4
     1       51-53 TLK

     2     106-108 TTK

     3     107-109 TKK

     4     111-113 TLR

 [3] PDOC00006 PS00006 CK2_PHOSPHO_SITE
Casein kinase II phosphorylation site

Number of matches: 4
     1       51-54 TLKD

     2       76-79 SGMD

     3     139-142 SYDE

     4     212-215 SKLE

 [4] PDOC00008 PS00008 MYRISTYL
N-myristoylation site

Number of matches: 4
     1      73-78 GIASGM
```

FIG.2A

2	77-82 GMDKTV
3	150-155 GIVLCE
4	158-163 GQVYAD

Membrane spanning structure and domains:

Helix	Begin	End	Score	Certainty
1	142	162	0.872	Putative
2	184	204	0.652	Putative

BLAST Alignment to Top Hit:

```
>CRA|1000682328847 /altid=gi|8051618 /def=ref|NP_057952.1| LIM
          domain kinase 2 isoform 2b [Homo sapiens] /org=Homo
          sapiens /taxon=9606 /dataset=nraa /length=617
         Length = 617

 Score =  485 bits (1235), Expect = e-136
 Identities = 241/265 (90%), Positives = 241/265 (90%), Gaps = 22/265 (8%)

Query: 13  LLPVKVMRSLDHPNVLKFIGVLYKDKKLNLLTEYIEGGTLKDFLRSMDPFPWQQKVRFAK 72
            L  VKVMRSLDHPNVLKFIGVLYKDKKLNLLTEYIEGGTLKDFLRSMDPFPWQQKVRFAK
Sbjct: 353 LTEVKVMRSLDHPNVLKFIGVLYKDKKLNLLTEYIEGGTLKDFLRSMDPFPWQQKVRFAK 412

Query: 73  GIASGM----------------------DKTVVVADFGLSRLIVEERKRAPMEKATTKKR 110
            GIASGM                      DKTVVVADFGLSRLIVEERKRAPMEKATTKKR
Sbjct: 413 GIASGMAYLHSMCIIHRDLNSHNCLIKLDKTVVVADFGLSRLIVEERKRAPMEKATTKKR 472

Query: 111 TLRKNDRKKRYTVVGNPYWMAPEMLNGKSYDETVDIFSFGIVLCEIIGQVYADPDCLPRT 170
            TLRKNDRKKRYTVVGNPYWMAPEMLNGKSYDETVDIFSFGIVLCEIIGQVYADPDCLPRT
Sbjct: 473 TLRKNDRKKRYTVVGNPYWMAPEMLNGKSYDETVDIFSFGIVLCEIIGQVYADPDCLPRT 532

Query: 171 LDFGLNVKLFWEKFVPTDCPPAFFPLAAICCRLEPESRPAFSKLEDSFEALSLYLGELGI 230
            LDFGLNVKLFWEKFVPTDCPPAFFPLAAICCRLEPESRPAFSKLEDSFEALSLYLGELGI
Sbjct: 533 LDFGLNVKLFWEKFVPTDCPPAFFPLAAICCRLEPESRPAFSKLEDSFEALSLYLGELGI 592

Query: 231 PLPAELEELDHTVSMQYGLTRDSPP 255
            PLPAELEELDHTVSMQYGLTRDSPP
Sbjct: 593 PLPAELEELDHTVSMQYGLTRDSPP 617  (SEQ ID NO:4)
```

Hmmer search results (Pfam):

Model	Description	Score	E-value	N
PF00069	Eukaryotic protein kinase domain	100.1	1.1e-26	2
CE00031	CE00031 VEGFR	4.9	0.14	1
CE00204	CE00204 FIBROBLAST_GROWTH_RECEPTOR	4.7	1	1
CE00359	E00359 bone_morphogenetic_protein_receptor	1.8	7.9	1
CE00022	CE00022 MAGUK_subfamily_d	1.5	2.5	1
CE00287	CE00287 PTK_Eph_orphan_receptor	-48.4	3.8e-05	1
CE00292	CE00292 PTK_membrane_span	-61.8	2.1e-05	1

FIG.2B

```
CE00291  CE00291 PTK_fgf_receptor                      -113.0     0.027   1
CE00286  E00286 PTK_EGF_receptor                       -125.1     0.0021  1
CE00290  CE00290 PTK_Trk_family                        -151.3   6.5e-05   1
CE00288  CE00288  PTK_Insulin_receptor                 -210.4     0.014   1

Parsed for domains:
Model    Domain  seq-f seq-t     hmm-f hmm-t       score  E-value
PF00069    1/2      16    79 ..     41   105 ..    52.1  2.3e-13
CE00022    1/1     124   153 ..    187   216 ..     1.5      2.5
PF00069    2/2      81   156 ..    129   182 ..    48.0  3.1e-12
CE00031    1/1     129   156 ..   1114  1141 ..     4.9     0.14
CE00204    1/1     129   156 ..    705   732 ..     4.7        1
CE00359    1/1      79   157 ..    287   356 ..     1.8      7.9
CE00290    1/1       9   218 ..      1   282 []  -151.3  6.5e-05
CE00287    1/1       1   218 [.      1   260 []   -48.4  3.8e-05
CE00291    1/1       1   218 [.      1   285 []  -113.0    0.027
CE00292    1/1       1   218 [.      1   288 []   -61.8  2.1e-05
CE00288    1/1       1   218 [.      1   269 []  -210.4    0.014
CE00286    1/1       6   218 ..      1   263 []  -125.1   0.0021
```

FIG.2C

```
   1 TCATCCTTGC GCAGGGGCCA TGCTAACCTT CTGTGTCTCA GTCCAATTTT
  51 AATGTATGTG CTGCTGAAGC GAGAGTACCA GAGGTTTTTT TGATGGCAGT
 101 GACTTGAACT TATTTAAAAG ATAAGGAGGA GCCAGTGAGG GAGAGGGGTG
 151 CTGTAAAGAT AACTAAAAGT GCACTTCTTC TAAGAAGTAA GATGGAATGG
 201 GATCCAGAAC AGGGGTGTCA TACCGAGTAG CCCAGCCTTT GTTCCGTGGA
 251 CACTGGGGAG TCTAACCCAG AGCTGAGATA GCTTGCAGTG TGGATGAGCC
 301 AGCTGAGTAC AGCAGATAGG GAAAAGAAGC CAAAAATCTG AAGTAGGGCT
 351 GGGGTGAAGG ACAGGGAAGG GCTAGAGAGA CATTTGGAAA GTGAAACCAG
 401 GTGGATATGA GAGGAGAGAG TAGAGGGTCT TGATTTCGGG TCTTTCATGC
 451 TTAACCCAAA GCAGGTACTA AAGTATGTGT TGATTGAATG TCTTTGGGTT
 501 TCTCAAGACT GGAGAAAGCA GGGCAAGCTC TGGAGGGTAT GGCAATAACA
 551 AGTTATCTTG AATATCCTCA TGGTGGAAAG TCCTGATCCT GTTTGAATTT
 601 TGGAAATAGA AATCATTCAG AGCCAAGAGA TTGAATTGTT GAGTAAGTGG
 651 GTGGTCAGGT TACAGACTTA ATTTTGGGTT AAAAAGTAAA AACAAGAAAC
 701 AAGGTGTGGC TCTAAAATAA TGAGATGTGC TGGGGGTGGG GCATGGCAGC
 751 TCATAAACTG ACCCTGAAAG CTCTTACATG TAAGAGTTCC AAAAATATTT
 801 CCAAAACTTG GAAGATTCAT TTGGATGTTT GTGTTCATTA AAATCTCTCA
 851 CTAATTCATT GTCTTGTCCA CTGTCCGTAA CCCAACCTGG GATTGGTTTG
 901 AGTGAGTCTC TCAGACTTTC TGCCTTGGAG TTTGTGAGAG AGATGGCATA
 951 CTCTGTGACC ACTGTCACCC TAAAACCAAA AAGGCCCCTC TTGACAAGGA
1001 GTCTGAGGAT TTTAGACCCA GGAAGAATGA GTGATGGGCA TATATATATC
1051 CTATTACTGA GGCATGAGAA GAGTGGAATG GGTGGGTTGA GGTGGTGTTT
1101 TAAGGCCTCT TGCCAGCTTG TTTAACTCTT CTCTGGGGAA CGAGGGGGAC
1151 AACTGTGTAC ATTGGCTGCT CCAGAATGAT GTTGAGCAAT CTTGAAGTGC
1201 CAGGAGCTGT GCTTTGTCTA TTCATGGCCC CTGTGCCTGT GAAACAGGGT
1251 TCGGTGACTG TCACTGTGCC TGTGGCAGTC TGTAGTTACC CAGAGAGAAC
1301 AAAGCTGCAT ACACAGAGCG CACAAGGGAG TCTTGTAACA ACCTTGTCCT
1351 GCTTTCTAGG GCTGAGTCAG GTACCACAGC TTGATCTCAG CTGTCCTCTT
1401 TATTTCAAGA AGTTGACATC TGAGCCATAC CAGGAGTATT GTATTTTGTT
1451 TGAGGCCTCT CTTTTTGGAG GAACATGGAC CGACTCTGTG CTTTTGTCTA
1501 TGCTGGTCTC TGAGCTCACA CAACCCTTCA CCCTCCTTTC TCAGCCAGTG
1551 ATAGGTAAGT CTTCCCTATC TTGCAAGGCT CAGCTCAAGT GTCAGCTTCC
1601 TCTACAAAGA CTTTCCTGGT TCCCCTCATT GGAGTGAACA AGAGTTGACA
1651 TGGTAGAATG GAAAGAGCAG AAGCTTTAGA ATGAGCCAGA CCTGAGTATG
1701 AATGCTAGAT CCACCACTTA GCTAGTCAAC CCTGCCCCCT GCCTCAAGTT
1751 TTAATTTTCC TATCCATTAA GTGAATATAA TAATACCTGT GTCACAGGAT
1801 TATTTTGAGA ATTAAATGAG ATTAGGTCTA TGAAAGCACC TAGCAGAGTT
1851 CTTGGCATAT AGGAGGCATT CATTAAATAT TTGTTCTTCC CCTTTTATAC
1901 CCATTACTTT TCTTTTTCTG AACTAAAATA ATACTTGGTT CTATCTCTGA
1951 AATAACATCC AAGTGAAAAA TCAACAACAT GAAAGAGCAG TTCTTTTCCA
2001 GTGGATTTGC TTCTTAAGGA GCAGAGATTA TGTAATCTAA CAGCCTCCAA
2051 CATACAAAGA GCTTTGTATC TAGAACAGGG GTCCCCAGCC CCTGGACCGC
2101 CAACTGGTAC GGGTCTGTAG CCTGTTAGGA ACCAGGCTGC ACAGCAGGAG
2151 GTGAGCGGCG GGCCAGTGAG CATTGCTGCC TGAGCTCTGC CTCCTGTCAG
2201 ATCAGTGGTG GCATTAGATT CTCATAGGAG TGTGAACCCT ATTGTGAACT
2251 GCACATGCAA GGGATCTGGG TTGCATGCTC CTTATGAGAA TCTCACTAAT
2301 GGCTGATGAT CTGAGTTGGA ACAGTTTGAT ACCAAAACCA TCCCCCCGCC
2351 CCCCAACCCC CAGCCTAGGG TCCGTGGAAA AATTGGCCCC TGGTGCCAAA
2401 AAGGTTGAGG ACTGCTGATC TAGAGGACCA ATTTATTCAA TGTTGGTTGA
2451 GTAAATGAGC TCTTGGATTA GGTGATGGAA AAATCTGAAA AAACAGGGCT
```

FIG.3-1

```
2501 TTTGAGGAAT AGGAAAAGGC AGTAACATGT TTAACCCAGA GAGAAGTTTC
2551 TGGCTGTTGG CTGGGAATAG TCATAGGAAG GGCTGACACT GAAAAGAAGG
2601 AGATTGTGTT CGTTTCTTCT TCTCAGAGCT ATAAGCAAAG GCTGAAAGTT
2651 CTAGAAAAAG GCAAGTTTTG TTTCAGTAGA AAAAAGGATA ATCAGAACCA
2701 TTTTTAGAAA ATGGAATGAG ACTACTTTTG AGGCCATGAG TTCCTTGTCC
2751 CTGGAGAGAT GAGCAGAGGT TGGACAAGTG CTTACCAGAG ATCTTGTGGA
2801 GGCAGAAACT GTGCATCTAG CAGAGCATTG GCCTAACCCT TTCAAATGAG
2851 ATGCTGTTAA CTCAGTCTTA TTCTACATGG TAGGAATCCT GTCCCTTTGC
2901 CTCCTGCTAC TTTGGGCCTC TCAACCTCTT GGTTTTGTGT GCAGGTGAAG
2951 ATGTCTGGAG GTGTCCAGGC TGTGGGGACC ACATTGCTCC AAGCCAGATA
3001 TGGTACAGGA CTGTCAACGA AACCTGGCAC GGCTCTTGCT TCCGGTAGGT
3051 GGGCCTATCC TCCCATCTTT ACCAGTGTAC TATGGGCCAA GCACTATTTC
3101 ATGTTCTGAT GGAAAACACA GAAACAAGCT TCTGAGTTGA GAATTTCAAT
3151 CTTAGGGTGG GGAAAGGAAT GTACCAAGGA AGAGCTCATG ACCAAACCTC
3201 AAGTGTGGCC CCCCTGAACC CAGGTTAAAT TGGAAGAGCC ATAAATGGGC
3251 CAGCTGGAGG CAGGGTGGGG GGATGAGAGG AGCCCTTTCC AGGGTTGTCC
3301 CATATCCCTC ACTTTATGGG TGAGGAAACT GAGGCCCAGG AAGAGTGACT
3351 TTCCTGTGGC TGCACTACAG ATTATGCAGG TACTTCAAGA GTTGTTTGTA
3401 TTCTTATTTT ATTTTATTTT ATTTTATTTT ATTTTATTTT ATTTTATGAG
3451 AGGGATTCTT GCTGTTGCCC AGGCTGGAGT GCAGTGGTGC AATCTCGGCT
3501 CACTGCAATC TCTGCCTGCT GGGTTCAAGT GATTTTTCTG CCTTAGCTTC
3551 CTGAGTAGCT GAGATGACAG GCACCTGCCA CCATGCGCAG CTAATTTTTG
3601 TATTTTAGTG GAGACGGGGG TTTCAACATG TTGGTCAGGC TGGTCTTGAA
3651 CTCCTGACCT CAAATGATGC ACCCACCTCG ACCTCCCAAA GTGCTGGAAT
3701 TACAGGCGTG AACCACTGTG CCCAGCCAAG AGTTGTTTTT AGTGTGGTTG
3751 GCAGAGCCAG CTCTTCCTTC ACCACAGGAT GCCTCCCTAG GTTCCTACTT
3801 TTTGTTACTA GCTTTTATTA TAGCTATATT ATTATTATTA TTATTATTAT
3851 TATTATTATT ATTATTGAGA CAGAGTCTCG CTCTGTCGCC CAGGCTGGTG
3901 TACAGTGGTG CGATCCCGGG CTCACTGCAA CCTCTGCCTC CCGAGTTCAA
3951 GCAGTTCTCC TGCCTCAGCC CCCCGAGTAG GTGGGACTAC AGGCGCCTGC
4001 CACCACACCC GGCTAATTTT TGTATTTTTA GTAGAGACGG GGTTTCACCT
4051 TGTTGACCAG GCTGGTCTGG AGCTCCTGAC CTCAGGTAAG TGCTAGAATC
4101 ACAGGCGTGA ACCACTGCGC CCAGCCAAGA GTTGTTTTTA GTGTGGTTGG
4151 CAGAGCCAGC TCTTCCTCAC CACAGGTTGC CTCCCTAGGT TCCTACTTTT
4201 TGTTACTAGC TTTATTATAG CTACATTATT ATTATTATTG TTATTATTAT
4251 TGAGACAGAG TCTCGCTCTG TCGCCCAGGC TGGTGTACAG TGATGTGATC
4301 TTGGCTCACT GCAACCTCTG CCCCCCGAGT TCAAGCAATT CTCCTGCTTC
4351 AGCCCCCCTA GTAGGTGGGA CTCCAGGCAC CTGCCACCAC GCCCAGCTAA
4401 TTTTTGTATT TTTAGTAGAG GCGGGGTTTC ACCTTGTTGG CCAGGCTGGT
4451 CTCAAACTCC TGACCTCAGG TGATCCGCCT GCCTCGGCCT CCCAAAATGT
4501 TGGGATTACA GGCATGAGCC ACCGCGCCCT GCCTATAGCT ACATTATTTT
4551 TGTAGGCAGC TCAGTTTCTT AAAAATTATA CAGACTTCAA ATCAGATTTG
4601 TTCCTGCTGT CTGAGGCTCA GTTTCTTCAT CTGGAAAATG GATGGTAATA
4651 ATCTTGTTGA GATTGAATGA AATAATATAT GCAGTGTATC CAGTACATGG
4701 TAGACACCCA GTGAATGGTT ATTCCTTCCT CCCATCGGAT TGGAATTCTC
4751 AAGGGTGGGA ACTTGTCTTT ATATTCTTCA CAACGTAAAA TAGTTGAAAT
4801 TTGTTGGTGG AAAGAAGAGC AGTCCACTCC AGAGGCTGGA TGGGCATGCC
4851 TGGCCCCCAA GGTCTGAAGT GGTAGGGCTG TGCCTATATC CTGAGAATGA
4901 GATAGACTAG GCAGGCACCT TGTGCTGTAG ATTCCAGCTC CTGCACATAG
4951 CTCTTGTTGT AAAACATCCC TGTGCTTATA CCAAGTAATT GAGTTGACCT
```

FIG.3-2

```
5001 TTAAACACTT GCCTCTTCCC TGGGAACCAT ATAGGGGATT GGCCTGGAGA
5051 CGTCTGGCCT CTGGAAGAGT TGGAAAGCAG CCATCATTAT TATCCTTTCC
5101 TTTCAGCTAT AACTCAGAGC TCTCAAGTCT TTTCTGTGGA TCTTATTGCC
5151 TTGGTTCTTG CCCCTTTTAC TCCCAGGGAA GTTGATTCTG TCTTTTCTGT
5201 TCCATTTAGT ATGACAGGAG CAGAGAATGT CAGAGCTGTA AGGGACCTTA
5251 TAGTTAAAGC CTTTGGCTGG TCCTTTCATT TTATAGCTGG GACTAATAAG
5301 TAACGTCAAA ACCCAATGAG TTCACAGATT GGGTCTCGCC TTGGCATGTA
5351 ACCCATATGT TCATATTCTT GCTGTTTTCC TATGTGTATG AATATTTTCT
5401 ATCCAAAATA AGCAGGACAG GGTAGAGCAA GTTAATCTTT GGAATTTCTG
5451 GATTCTCTTA GAGCTAAAAA ACTTCAGAAC TAGAAGAAAC CACCCACTAT
5501 ATGGTATAAC CCATTCATAT CACAGATGAG GCCTGAAACC AAAAAGACTT
5551 GCTCAGGCCA TGGATGACAA GAGCTGGCCC TAGCACTGAA CTCTTGGGTC
5601 ATTTGTAGGT CTAGTCAGAT GCTAGCTTGT TAGCTCTGTG CGTGCGTGTG
5651 TGTGTGTGTG TGTGTGTGTG TGTGTGAGAT AGAGACAGAA AGATAACATA
5701 TGTACACAAA TACATAAAGA GGAAGTAGAC ACGTTAGCAT GGTAGATAAG
5751 AGTACAGGCA GGCCAGGCGT GGTGGCTCAC GCCTGTAATC CCAGCACTTT
5801 GGGAGGCCAA GGCAGGTGGA TCACCTGAGG TCAGGAATTC GAGACCAGCC
5851 TGACCAACAT GGTGAAACCC CATCTCTACT AAATACAGAA AAAAATTAGC
5901 TTGGCATGGT GGCACATGCC TGTAATCCCA GCTACTTGGG AAGCTGAAGC
5951 AGGAGAATCG CTTGAATCCG GGAAGCAGAA GTTGCAGTGA GCCGAGATTG
6001 TGCCATTACA GTCTAGCCTG GGCAACAAGA GGGAAACTCC ATCGCAAAAA
6051 AACAACCACC ACCAAGAGTA CAGGCTATGG AATGAGACTA TGGTTTTAAA
6101 TCCTGGCTTT GCAATTTATT AACTAGCCTT AAGTGACTTC CCTGAGCTTC
6151 AGGCACCAAT CTGTAAAATG AGGATAAGAA TATTACTCAT GCCACATGGT
6201 TGTTAGGGAG GATTAAATGT GATAACCTAT ATAAAGTGGC TAGCATAGCA
6251 TCTGACATAT AGAAAACTCT TAATAGGGCC GGACGTGGTG GCTTATGCCT
6301 GTAATCCTAG CACTCTGGGA GGCCGAGGCA GAAGGATCGC TTGAGCCCAT
6351 GAGCCCAGGA GTTTGAGACC AGCCTGGCCA ACATGGCAAA ACTCCACCTC
6401 TACAAAAAAT ACAAAAATAT TAGCCAGGCG TGATGGCACA CACCTGTAGT
6451 CCCAGCTACT TGGGAAGCTG AGGAGCGATG ATTACCTGAG CCCAGGGATA
6501 TCAAGGCTGT AGTGAGCTGT GATCATGCCA CTGTACTCCA TCCAGCTGGG
6551 GGACAGAGTG AAACCCCTGT CTCAAAACAA AACAAATGAA AAAAAAAACC
6601 CTTAATAATC AGTAACTGTC ACTTTATATT ATGTTGTGAG TGTGTGTCTA
6651 TATACACCTA TATGTATACA TTTCTCTTAT TACACATTCA TTGGTGATCT
6701 GATGTGGAGC CCCAGGGATT AAGGGCAACT TTGAACTACC CTGACACAAT
6751 CAAGCCAAAT ATCATTCCCG TGGAGGAAGT AGAGTATCTA GGTTCTGTCT
6801 CCTAGTTGCA GCTTTACCTT GAGGACAGAG ACTCTAATCC AGCTGTGCTG
6851 AAGGAGCACA TCTCCTGACT TCTGAGCTTT CCCCTGGTAA ATTCAAACTG
6901 GATGTCACGG CGCCCTCAGA TAGAGCCTGG TAATTTGCCC TGGGGAGAGT
6951 GACTGTCTTT TGGATCTAAT TTGACTTTTG CCCCAGTTGG AGGAAAATCT
7001 TCAGGGCTAG GAAGGATTGT ATTTGTCTGA CCCCAGAGAT AACCTGGGTT
7051 TTGAGGAACA TGGGGCATCA ACCTGAATGG TCTTGTAAGA TCTCTCCCAC
7101 GCCAGCTTGC CAGTGTTTCT CTGATGAATT TAGAGTACCT GAGTAGTGCA
7151 GGCCTGCTGG GAGGAGGACT CTCCCTCTGT GCTACTCAGA GAAATTCATT
7201 CTTCAAGGCC CCCTTCCAGC CTTGCTCTTA CCCAGCTGGG CTACAGTTAC
7251 AATAAAGGAA ATGACTTTTC TTCTCCCCTT CCCCCAGTAC CTTTGTTTTC
7301 CTAGTCACAG GGTGGGGCTG GATATTGAAT GGAGAAATTG CTGGGGTCCA
7351 TCCTAAACTC CTCCCCTCAT CTCTCCCTTA CATTACCCCA TTCTTCTGTC
7401 TGCAGCCACA TCCATAATCC TGCCTCTGTT AGCCTTCCGA CAGACCCTCA
7451 GGTGCCCAGG ACAACAGGAA GCTACTTAAA GCTGGAACCT CAGACTGTGC
```

FIG.3-3

```
7501 AATGGAGGCC AGTGACAAAA CTGAAAGTAG CTCTGTCAGT AATTGTGCTG
7551 GTGCGATTAG GCAGCTGGCC AGAATCTTTT GGATCTCCTG GACATATGGC
7601 TGACTAGTCC TCCCAAGCCT TCCCAACAGG CCTCTTTTTT TTCCTTTTTT
7651 TCTTTTCTTT TTTTTCTTTC TTTCTTTCTT TCTTTTTTTT TTTTTTTTAG
7701 GCTAGTGAAG TGAAATTGTG GGAGTGGAAA AGGAACAAAG AAATCGGTAA
7751 CTGGTAGTGA TCAATTACTT GTAAACACTA TTGTACTTGG ACCAGCCCAG
7801 TAGGCCTTTT TTAAAACTCT GAGTTACCTC TCTTTCCTTT CCTTGAGCAG
7851 TGCCATTAAT TCTGTATCTG GGGCAATCCT TTCTGATGTT CTCTGGACCT
7901 GGCTCTCTCT CCTTAGGAGA GGCCAGGAGA GTAGCCAGAG AGCATGTCAT
7951 TTGTAGCTGA GGTTAAAGTG TGGAGCTATC AATGGTGACC TGGCCTCTTG
8001 GCATGTTAGC AAGCCAGAGG ACCTTGACAA CTTTTTTGAT GATTGTCCGT
8051 TCACCCTGAT CAAAGGTGTT TGGCTTAGGA GGAGGGAAGA AAAGCTACCC
8101 CTATTAGTCT TGATGGCCCC AGCGTGGGTC TCTATTGCTT GACCTGGTTC
8151 CTAGCAGCAT TATCAGAAGG AAAATCCACC GCTCTTAAGG CTCCTGGGAA
8201 CTTTCAGGAC TTCCTTTCTC AGGATTGCAA ACATAAGACT ATTTGAGCTT
8251 TCACTTTTGA AAAGCGGTTA CTAATACCTA TACTCTGGGA AAGGGCTAAT
8301 GCAGATAGAA GACTGTGGTC ACTGCATCAG GCAACAGACC ATTTCCGCTA
8351 AATTTAGTGA CTCCAGGAAG GCCAGTGAAG AAATAACACA CGTAGCAACC
8401 AGAGACTGTG TTGTAATATG TTGGCTGACA GCAGGGTACT TTCTGTGATG
8451 CTGAAAGCCA CATTCATTTT CTCTCCCCTC ATCCCCATCT AAGCAAGCCT
8501 GGTAGAATCA TAATTACAGT AATAGGTACC ACTTATTGAG TACTCTGTGC
8551 CAGACACCCT CCTGAGCATA CGACATGCAT AGCACATTTA ATCCTTACAA
8601 TGACTTAATA AAATGTAGTA CTAGTCTTAC CTACTTCGAG AATAGGGAAA
8651 TGGAGGTTAC TTGTTTAAAG TCACAGAGCT AATAGGTAGC ATAGCTGAGA
8701 TTTGAACTCA GGCATTCTTA CTCCTTGCCT GCAAGAGTCT CTTGGCATTC
8751 TTGAATGCAA GCATATTTCT TAACCTCACT GAGGCTCAGT TTCCTCTTAT
8801 ATAATATGGG GTAAAGAGCC CTCACCCTGC CTGCCACACA CTGGTAGTGT
8851 CAGATAACAT TGAAGGGTGT TAGTTTAAAG GCTTCATGGA CTCTATAATG
8901 TCAACAAAAG TGCTGTTAAC TTTCTTCTGG GTCTCAGGCT CCTGATGTAG
8951 AGTCAGTGGA GCAACCCTGC CATCTGCTGT TATGCTGTTG ATGTTGCTGC
9001 CACACTTACT AACCTAAACC TTTGATTCTG GCTGTGGCCT TCTCCAGAAG
9051 GTGTTTACTC ATTTGTCCAG TTTATCTTTT AGGAAACAGC CAGCCCGTAG
9101 ATCATTAAGG CTGGCTATTG GACAGGGGGC TGGGGCCTGC CTGACAGAGG
9151 AAGGAAGGGC AGACATCTGG TTCTTCCTCT GCCCCTACAA GAGACTCCAG
9201 CCTGACCACA GAGTGGTACT CCTAGGATGT AGCAGCAGCA TATGAGCTTG
9251 AATGTGCCTT AATCCTGCTC TTTACTTTGA GAAGAGAGAA CTAAGGACCC
9301 ACAGATGTTT CACAGCTTCT ATAGGAGGCA GAGGTAGAAA AATGGAGAGA
9351 GATGAGGCCA GAGATAGATA ACTGATATTA ATTAAACGTT GTATTAAGAA
9401 CCTCACTTAG ATTATCTGAT TCAATCTTCA TAATAACCCT GCAACCCCCA
9451 CCTTTTTTTG AGAACAGGGT CTTGCTCTGT TGTCCAGGCT ACAGTGCACT
9501 GGTACAATCA TAGTTCACTG CAGTGTCAAC CTCCTGAGCT CAAGCAATCC
9551 TCCCACCTCA GCCTTGCAAG CAGCTTGGAC TACAGGCGTG CCACCACACC
9601 TTGCCATTTT TTTTTATTTT AAGTAGAAAC AAGGTCTTAT TAATACTATG
9651 TTGCCCAGGC TGGTCTTGAA CTCCAGCGAT CCTCCTGCCC CAGCCTCCCA
9701 AAGTGCTTGG GATTACGGAA GTAAGCCACT GTGCCTGGCC AGTGCAACCC
9751 CCATTTTATA CTAAAACAGG AAGGCCCAGA AAGGTTTGGA GTAACTTGTC
9801 CAGGGTCACA CAGATGATAT TTGAACTCAG GTCTCCCTGG CTCCCAAGAG
9851 AGTCTGCTTT CCACTAGGAC TCCCAGGAGA AAAAAAAAAA AAAAAACAGT
9901 AGACTTGGAG ACAGAAAATC TGATTTGAGT CTTAGTTGAG CTAGGCTAAC
9951 TGTGTAACTG TGGGCAAGTT CCTTAGCCCC TGTGAGCCTC AGTTTCTTAT
```

FIG.3-4

```
10001 CTGTAAAATG TCATAAAAGA AATCCATCTC ATGGAGTAGT TGTGATGATC
10051 AAGGACTCTG AAAACATTAG AATGGTTTAA TGTGAAGGAT TAGCAGCAGC
10101 ACATGGCAAC ATTGTGCATC TTATATTAAC TATCCAAATA TATCAAGCGT
10151 CATTTGCTAT ATATAAAAGT CATCAAATTA GGCACTGTGG GGGATACGGA
10201 GTTGGCATAC TAGCCTGGCC TCTTAATTAA TTCATTAATT AGCTTATTTA
10251 TTTTTGAGAT AGGTCTTGCT CTATTGCCCA GGCTGGAGTG CAGTGGCATG
10301 ATGATAGCTT ACTATAGCCT CAATCTCCCA GGCTTAAACA ATCCTCCTGA
10351 GTAGCTGGGA CTACAGGCAC ACACTACCAT GCCCAGCTAA TTTTTTTTTA
10401 ATTTTTTGTA GAGACAGGGT CTTGCTCTGT TGCCCAGGCT GGTCTCAAAC
10451 TCCTGGGCTC GAGATCCTCC CACCTGGGCC TCACAAAGTG TTGGGATTAC
10501 AGGTATGAGC CACGGCACCT GGCCTGGTCT CTTAACTGGT TCCCTAAGAC
10551 AGCTGGAAAT AGAGAATGTC ATGGAGCATT CCTAACCATG GGCTCCAGCC
10601 TGGCTTTCAT TCTGTTTCTC CCCTGAAACA ACATTCCTTT AGTAATATTC
10651 CGAATAACAG CTTCATCAGT CTGTCTACCG ACCACTCTTC AGGCTTCATC
10701 TTATATGACC TCCCAAACTG CACTAAGGGT TGTATTAGAG AAAAGTGGAT
10751 AAAGTTCGGA GTCAGGCTGC TTGAGCTTAA ATGCCAGCTT CACTTACCAG
10801 CCACCTGACC ATGAGTCAGC TGCTTAACCA TTCTTTGCCA CAGTTTCCTT
10851 GTCTATGAAA AGGGAAATGG CTCCCACCTC AAAAAGTTGT TAACATTAAA
10901 TTCAATCATG TATTCAAAGT CCTGAGCAGA ATGTCTGGCC ATGACTGGGA
10951 CTTAACAGAT GTTAGCATTT ATTATTAGTA TCTGTCAGTC TTGAAATGTT
11001 CTCTTCCCTT GGCTTTCATG ACATTCCACA CTCTCCTGGT TTTCTCTTAC
11051 CTCTCTGGTA ATACCTGTTT GCTTATCCTT CTTTGTCCAG CTCTGGGATG
11101 TTACCATTCC TTCAGGCGTG CTGTTTTCTC CTTAGGCAGT CTTACACACA
11151 CTCATGACTT CCTTCCATTG TCCTCCACAC ACTGATGACC CTAAAATCAG
11201 TATCTCCAGC CTAAACCTTT CCACTGAGTT CTAGACCCAT ATGTTGTACT
11251 ATCAACCTGG CTTGTCCATT TGAATGTCTT CCAGGCACTT CAGACTCTCT
11301 TCTCTAGACT TTGCTGGACT TTCACTCTTC CCCCTAAAAC TGGCTCCTCT
11351 TCCACTGAAA CATGTATGTC ATTGAGAGGC ACCACCATCC ACCCAGTGCC
11401 TAAGCCAGAA ACCTAGGAAT CCTTGATACC TGTTCTCTCT CATCCTGCAT
11451 ATCCAAGCCT ATCAGTTTTA TCTCTAAATT ATATTTTGGT AGGTTTACTT
11501 CTTTCCTTTT CTCCCACCAC CACCCTGCTC CAAGCTACCA TCATCTCACC
11551 TGGATGTCTG CAATAGCCTC ATCTCCCACA GCCACTCTGC ACCCCCTAAT
11601 CTGTTCTCTA TAGAGCAGTT GGAAGGAGTG ATTTTTGTTG TTTGTTTTGT
11651 TTTGTTTTAG ACAGAGTCTC ACTCTGTTCC CCAAGGCTGG AGTGCAGTGG
11701 CACAATTTCG GCTCACTGCA ACTTCTGCCT CCCGGGTTTA AGCAATTCTC
11751 CTGCCTCAGC CTCCCAAGTA GCTGGGATTA AGGCACCGGC CCCCATACCC
11801 AGCTAATTTT TATATTTTTA GTAGAGATGG GGTTTTGCCA TGTTGGCCAA
11851 GCTAGTCTCG AACTCCTGAC CTCAAGTGAT CCACCTGCCT CGGCCTCCCA
11901 AAGTGCTGGG ATTACAGGTG TGAGCCACTG CACCTGGCTG GAAGGAGTGA
11951 TCTTAAAAAA AAAAAAAACA AAAAAAAACT TGACTGTGTC ACTCTGTGTT
12001 GTCTCTCCTA CCTTGTATAC TTCCACAACT TCCCAGTGTT CTTGGATAAA
12051 GACCAAAATC CTTAACTTGG CCAGGCGCGG TGGCTCACAC CTATCATCTC
12101 AGCACTTTGG GAGGCCGAGG CAGGCAGATC ATGAAGTCAA GAGATTGAGA
12151 CCATCCTGGC CAACATGGTG AAACCCCATC TCTACTAAAA ATACAAAAAT
12201 TAGCTGGTCG TGGTGGCGTG TGCCTGTAGT CCCAGCTACT TGGGAGGCTG
12251 AGGCAGGAGA ATCACTTGAA CCTGGGAGGC AGAGGTTGCA GTGAGCCCAG
12301 ATCACGCCAC TGCACTCCAG CCTGGTGACA GAGTAAGACT CCATCTCAAA
12351 AAAAAAAAAA AAAAAAAAAA TTCCTTAATT TGGCCTACAG TAGAGCCCTC
12401 CGTAATGTGG CCTCTCTCCA CATCTCCACA ACCTCCTGCT CCCTGCACTT
12451 CAGCCTCACC TCTCTTCTGG ACAGGCCCTC CTTCTGACAA GGGCTTTGTT
```

FIG.3-5

```
12501 CATTCTGCTC CCTCTGCCTA GAATGCCCCC TTACTCTGTT CACTTAACTC
12551 CTGCTTATCG TTTAGATCTT TACCTGGATG GCTCAGAGAA ATATAGAAGT
12601 AATTCCTCAC CCTGAAAAAT AGGTTAGGTC CCTGTTTTAT GTTTTCATAG
12651 ACCTTTCCTT TGAGGCTTTT TTTAAAAAAG TAGTTTTAAT CTCACATTTA
12701 TTCATGTGAT CATCTCCTTA ATGATATCTT AAGACCTCTA ATAGAACAAT
12751 TTGGTCATGG ACTGTGGGGT TTTTGCCCCT CATTGTGTCA GCACTGAGCA
12801 TATTGTTGGC ATAGGAGGGA TATTTGTTGA ATGAATTGCT AGAGGTGGCC
12851 AAGAGATATG ATGTAAGTCA GGCTTTTCCC TGCCCTTCCC CTTCCCCTTC
12901 CCCACATCCT TCCTATAGCA GCCACCGTGG CTGCAGTTAC TGTAAATGGC
12951 AAGACGGAAT CAGTTCCGGA CATTGGGTTG TTTTAGAAAA TTGCCTGCAA
13001 GTGTCAGGGT GATAAGTTAA AGCTTTGTCT TTTGCCCTCA GAGGAGCTAT
13051 CCCATAGTGA GTAGAAGCCA GAGAAGCTGA CCCCAGGAGT CCTTCTTTCC
13101 AGCAGCAGGT CTTGAGCTGC ACTTCTCTGT AGCTACAATC CAGGCAGGAA
13151 CAAGCCCTAG GTACCTCCGG AGAGGAGGGC AAGAGAGGAA GAATGAGTTC
13201 AGCTACTCTA GCCACCAAAC TGATTATGAA TTGCCCTGAA ATCTGAAAAA
13251 TTTCAATTCC AATCGTAAGT TTGTTTTGTT TCATTTTGTT TTCTTAAATT
13301 GTATATTTGA AAGATGGCAT TAACTAAAGA TATATATTCA ATATAGAGTG
13351 GAAAAAATGG AATACTTGCA TAGTATCTTT TACTTATAGG TGATTTATGA
13401 TGGGGAGTGG GGTGGATAGG TTGGCAGTTC CCCCAAGAAG TTGGAAATGA
13451 AGTTTGTCCT CTGTGAGTTG AACTAATTAG ATCCACAAGT AATGAAAGCA
13501 GTATTGTGTT GTAGTTAAGA GCACACTCTA GAACCAGATT GCTTAGTTTC
13551 AAATCCTGGT TCTGCCTTTT ATTATCTGTG TACTTTGGGC AAGTTACTTG
13601 CCCTTTGTGT GCTTCATTTT TCTCATCTAG AAAATGGAGA GGCCAGGCGT
13651 AGTGGCTCAT GCCTATAATC CCAGCACTTT GGGAGGCCGA GGCGGGCAGA
13701 TCACCTGAGG TGAGAAGTTC AAGACCAGCC TGGCCAACAT GGTGAAACCC
13751 TGTCTCTACA AAAATACAAA AATTAGCCAG GCATGATGGC GGGTGCCTGT
13801 AATCCCAGCT ACCCAGGAGC CTGAGGCGGG AGAAACACTT GAACCTGGAA
13851 GGCAGAGGTT GTAGTGAGCC AGGATTGCAC CACTGCACTC CAGCCTGGGT
13901 GACAAGAGCT AGACTCAGTC TAAAAAAAAA AAAAAAAAAC AAACTGGAGA
13951 TACAGGCTGG GTGCAGGGCT TACACTTATA ATATCAGCAC TTTGGGAGGC
14001 CTAGGCGGGA GGATTGCTTG AACTCAGGAG TTTCAAGATC AGTCTGGGTA
14051 ACAGAGCAAG ACCTCATCCC CACAAAAAAT CAAAAATTTA GCCAGGCATG
14101 GTGGCTCATG CCTGTGGTCC CAGCTACTCA GGAGGCTGAG GCGAGAGGAT
14151 TGCTTGAGCC CAGGAGGTTG AGGCTGCAGT GAACCATGAC TGCACCACTA
14201 CATGCCAGCC TGGATGACAG AGCAAGACCC TATCTCAAAA AAAAAAAAAA
14251 AAAGAAACGA GCCAGGCGCG TTTGCTCACG CCAGTAATCC CAGCACTTTG
14301 GGAGGCCAAG GCAGGTGGAT CACTTGAGGT CAGGAGATCG AGACTAGCCT
14351 GGCCAACATG GTGAAACCCC ATCTCAACTG AAAATACAAA AATTAGCCAG
14401 GCATGGTGGC ATGCTCCTGT AGTCCCAGCT ACTCACTTGG AGGCTGAGGC
14451 ACGAGAATCG CTTGAACCCA GGAGGCGGAG GTTGCAGTGG GCCAACATCA
14501 TGTCACTGCA CTCCAGCCTG GGAGACAGAG CGAGACTCTG TCTCAATAAA
14551 TAAATAAACA TAAAATAAAA TAAAATAAAA TAAAATAAAA TAAAAAAATA
14601 TGGAGGCCAG CAGGCACGGT GGCTCACGCA TGTAATCCCA GCACTTTGGG
14651 AGGCCGAGGG GGGCGGATCA CAAGGTCAGG AGATCGAGAC CATCCTGGCT
14701 AACACAGTGA AACCGCGTCT CTACTAAAAA TACACAAAAT TAGCCAGGCA
14751 TGGTGGCAGG CACCTGTAGT CCCTGCTACT CAGGAGGCTG AGGCAGGAGA
14801 ATGGCGTGAA CCCGGGAGGC GGAGCTTGCA GTGAGCTGAG ATCGCGCCAC
14851 TGCAGTCCAG CCTGGGCGAC AGAGCAAGAC TCTGTCTCAA AAAAAAAAAA
14901 AAAAATGGAG GTTGGGCGCG GTGGCTCGCG CCTGTAATCC CAGCACTTTG
14951 GGAGGTCGAG GCGGGCGGAT CACCTGAGGT CAGGAGTTCC AGACCAGCCT
```

FIG.3-6

```
15001 GGCCAACATG GTGAAACCTT GTCTCTACTA AAATTACAAA AATTAGCCAG
15051 GCACGATGGC AGGCACCTGT AATCCCAGCT ACTTAGGAGA CTAAGGCAGG
15101 AGAATAGCTT GAACCTGGGA GATGGAGGTT GCAGTGTGCT GAGATCGCGC
15151 CACTGCCCTC CAGTAGAGTG AGATTCCGTC TCAAAAAAAA AAAAAAAGAA
15201 GAAATGGAGA TACAAACTTA CTACCTACCT CCTTACAACC TACCCTCACA
15251 GTATTACTGT GAATAAAAGT GTGTGTAGCA CTGGGAACAC TATTCACAGA
15301 GCACTCATGA ATGTTTGTTC TTTGTTATTA GTTACTAGAG AGGCAAATGT
15351 CTGCCAGGGC TGAATAATAT GTGTGAATTG GTGATTGTCG CACATATCTA
15401 AAGAAGTAGT TATTTTTTTC AATTAAAACT TAGTTTAAAA ACCAATATAA
15451 GGCCGAGCGC AGTGGCTCAC ACCTGTAATC CCAGCACTTT GGGAGGCCGA
15501 GGTGGGCAGA TCATTTGAGG TCAGGAGTTC GAGACTAGCC TGGCCAACAT
15551 GGTGAAACCC TGTCTCTGCT AAAAAAAAAA AAAAAGTACA AAAATTAGCC
15601 AGGCATGATG GCAGGTCCCT GTAATCCCAG CTACTTGGGA GGCCGAGGCA
15651 GGAGAATTGC TTGAACCCAG GAGGTGGAGG TTGTAGTGAG CCGAGTTTGT
15701 GCCACTGCAC TTCAGCCTGG GTGACAGAGG GAGACACTGT CTCAAAAAAA
15751 AAAAAAAAAA ACCAAAACCA ATATAATAAA TAAGTGGCCA GCAATGAAAC
15801 AGAAAGTGAA AAGTTAGTGA AGCAAAACTA GTACTGTATT CAGATAAAGA
15851 TGCTGAATCT AGATTTGGTC ACCAGAATAG GGTCCTTTGT GGCAACCTGG
15901 GCTAGTTTGG CTGACTCACC ACTGCCAGGA TGAAATTTCT TTCAGTGGCT
15951 ACTCATTTCC CTTTATTTTA AGTCCATGCT CACAGAGCAA CCTTCTGATG
16001 CCTAATTCAG CTTCCTGGGA TACTTAATAA CAGGAAGGGT CTGGAAGTAG
16051 TACCTGTATA GGGGATATGA GTGTTCTGAT TTTAATAGTC AATTCATAAG
16101 TGTACAGAGG GTTTGATAAA TGGTTAGGTC AGAACCATCA CAGAATGTCT
16151 ACACCTCTTT GGACATTAGG AAGGTCAAAA ACCTGAAAGG CCAAAAGCTA
16201 GGCCTAGATT AGGGTCATTC ACCAAGAAAA CATCAGCCTT GAAGAGTTCT
16251 CTGGGTGGTC CACCAGTCAA CCTTCCTTTG ATCACACCTC CTTCCTCGTT
16301 GCTTCTTTAA GCATTGACCT GTAATGGGTA TGGAATTTTT TGCTCACCTA
16351 ACTCCTTCCT TTTACAGAGG AAGAAGTTGA AGCCCAGAGA GATTTAATGG
16401 CTTGCCTAAG ATCACACGCA GATTTTCTGT TAACCAGGGT GATTTTTCAG
16451 GTGTTCCCTG CCAGACGAGG GCTTTTTTCC TTGAATTGCC TAGAGATTTC
16501 TTGAGATATC CGAAGCATTT TTCCCAGTGC AGCCTGGAGA AGGATGTCCC
16551 TGTCAACACA GCATTTGTTA CTCAATGTTA GACATTCAAT TTTCTAATTA
16601 GTATCATGGA GCAACAGTGG ATGATTATCT ATAAGGGGTT GCAATTCCAT
16651 GCTTATGTGC TTACAGCCCA TATAGACAAA TATCAGCTGT TAAAATGACA
16701 AGGCAGTAGA GATGTGGCCC CAGGACAAAG GCATACTCTG CTGTTAGTGA
16751 ACACTAGTTG GCCAGCAAAT TTCACATGGG CATATACACG GCCAACTGTA
16801 GACTTTAGGC ATTTATACCC ATTCAGAGAG CCAAACTGGC AACTAAAGAT
16851 CAGCATTCTC TTTGGCATTT CAGCTTTGCG TTCTGTTAAA AATCACTGCT
16901 TGCTTAAATA CCTCTGATAG CTCTTCACTG CCTGTAGGCA ACTCTTTAGC
16951 CTAGCAGACT TGGTCTTTAG TGCTCTGCCC CTACTCTCTT CCACCATTCT
17001 GGCCTCCTGT CTAATTGCTG CCCATATGTG CCATGCACTA GAGCTTACAG
17051 ACCTGCTCAG CGTTATATGA GCATACCATA CTCTTTATGC CTCAGTGCAT
17101 TTGCACATGT TGTTCCTTCA GGCCAGAATG CCTGTTACTG CCTGGCAATC
17151 AGCCTATTAG AGTCTGCCAA TACCATCCCA TCTTCTGTGG AGGAGCCCCC
17201 CGCCAAATCC ACCCATACCT CTCCCCACCA ATCAGAGACT TCTTCTCTCT
17251 TTGTTATTCT CTTCGTTATT CTCTTCATAC CTCAGTTATA TCCATTTCAG
17301 TATTTGTTTA CACATCTAGC ATCACTCTTA GAGTGTGAAA TTCTCCAAGT
17351 GTGGAGCCGT ATCTAGTTTG TCTTTGTATC CCAGAGCTTA GCAAAGTGCC
17401 TAGAATGTAG TGGGTGCTCA GAGTGTTTGC TGGGTGAATG ATGTATTTGT
17451 TGAACGACTC TTTGGACACT TGAATAAAGT CCATCCAGTA TGCACCATTA
```

FIG.3-7

```
17501 CCATCTCTTC GCTCTACAAT ATTCTTTTAG GCAAGAGCTT ATCTTTTGAG
17551 GTGATAAGAT AAGCTCAAAC TTATGTAGAC TAAGACCTCA GTCTGTAAAT
17601 GTCATCCCTA AGTCTTAAAC CATCAAAACC AGGGCCTCAA GGAATGGCAT
17651 GCCTTCTGCA ACTGTAGCAA CCTGCTGTGC TTATTTTGCC GTGTTTTTCA
17701 TTTTTCCCCC AAAAGCTAGA GTCCCTTCTC CCATGGGCAG TGCTGGAAGT
17751 GTGCTAACAA ATTCTTTCTC CATACTGCTT ACGATTACAA AAAAAACCCT
17801 CAGCATCTCA TGCCAGACTT GAGTTAAGGT TGTTTTCTTT TGTGTGTCAG
17851 CTGTATTCTG GTCATGACTT CCTGATGATG CCCTATAGAG ATTTTGCTGA
17901 GATCAGAGGG TGCTCCACTG CCATCAGTAG CACTGACTCT TGCAGAAGCA
17951 CCGTTTCTGA AGTTGGCTAA TGTCATCCCT CACGTTTGTT TGTTTGAAAT
18001 TTGTTTTAGT TCCAGAGATA GCACTTTCAT GGAATGACGC TATCTTCTAG
18051 AATCACTTTT TTTTTTTTTT TGAGTTGGAG TCTCGCTGTG TCGCCAGGCT
18101 GGAGTGCAGT GGCACAATCT CAGCTCACTG CAATCTCCAC CTTCCGGGTT
18151 CAAGTGATTC CCCTGCCTCA GCCTCCCGAG GAGCTGTTAC TACAGGCGCA
18201 CACCCCCACT CCTGGCTAAT TTTATGTGTT TTAGTAGAGA CGGGGTTTCA
18251 CCGTGTTGGC CAGGATGGTC TCGATCTCCT GACTTTGTGA TCTGCCTGCT
18301 TCAGCCTCCC AAAGTGCTGG GATTACAGGT GTGAGTCACC GCGCCTGGCC
18351 TAGAATCACC TTTTTATACC ATAACGTGAG CACCACTGCC GCGTCACCAA
18401 GGAAAGAGAG AGGCAGCTAC TGTGGGGTTA CAAATGGGTA AGAGTGGCAC
18451 CAGGAAGGTG AAAGTCTCTA CTTAGCCAAG GCTTAACAAA ATGTCAATCA
18501 CCAAACATTT ATTTATTAAG CTACGTTCAG GATAAGAAGA TGAACAAGCT
18551 ATCTGTACAT TCATTTTCTC GTTTGTAACA AGGTAATGAT AGTGATCTAT
18601 CCTGCCTGCC TCTGAGGGTT ATTGTGAGAA TAAAATGAAA TCAAGTGGAA
18651 AAGCACTTAG GAAAAAGAAA AGCATTGGTT TTCAATTGTT AGTGTGGATC
18701 AGAAACACTG GGGCTTGTTT AAAATGCAGA TTCTTAGCCC CAGTCTCAGC
18751 GATTCTGATT CTGTATATCT GAAGTGGGAC TCAGGAATCT TGATTTTCAA
18801 CAAGCTGACC AGAGGGTCCA ATGCTGCTAT TCCTTTAGTT ACACTTTCAG
18851 AAATATTACT GTAAATCAAA TGGCAAGAAT AAAATAGTTA TTTGAGGCAG
18901 TTTTAGTATG TTGGACCTGG AGTCCAAAGA CTTGGGTCAA ACTCCAGCTT
18951 TGTCAGTTCC TAGACCTGTG ACCTTAAACA GCAACCTTCT CTGTGAACCT
19001 TAGTTCCCTC AGGAACGGCT CTGGTCACCT CCTGCTGTAC TCCATTGATG
19051 ACTCACCACA TAAGGCTCCC TGGGAGTCCC CCAAACCTTT GCTCTCTTAA
19101 CTCCTTTTAC AGCCTCCTAC ATCTCCTGCA GGTGCTGTCT TCTCCTCCTT
19151 TTTCCAGGCC CTGCTCTGAC ACAGCATTCA TTCTCCTCTG GGAAGGGTTC
19201 CTTCAATGTG TCTCCAAGCA CATCACACCC AGGAAGGACC CTGTGGCCAT
19251 ATCTGTCTAT CACCAGATCA AACTACGTGA AGGCAGGCAC TAGGTACTGT
19301 CAGTGCCCAG CATAGGCCTG GCCCATACCA GGTGTCCACA GATGCCTAGT
19351 AAAGAAACCT ATGATTCAGG ACCCCCATGA TGAGCAACTA TAGCACTAGA
19401 ACAGTGATAA TAACTAATGT TTATAATGCA TCTTCAGTTT ACAGAGGGCT
19451 TTTGTACTCA TCATCTAGTT TAGTTCCTGC AACAACCTCT TGAGGAATAT
19501 AGCACAAGCA GGACAAGGGA AGCCCAGAGA TGTTAAATAA TTTATCCAAG
19551 TTTATGCTGC TGGGAAGGGC AGCACTGAAA TTAAAAGAAA AGTTTTCTGA
19601 GCTCAAATCC CATGCCCTTT CCTCAATGTG AGCTCTAGCA AGGTATTCAG
19651 GAATCCTGCC TCTACAGTTC AGAGCCTCAA ATTGCTGGGT ATGTTGAGTT
19701 CTTGTATCTG ATTTTTCTAG ATTTCCTGCC CACATTCTTA CTGTCTGGAT
19751 ATCAGGAAAG AGTTTATCAA ATGCCTGTGG AAATCCAAGA TAAGGTCTCA
19801 TGATGAGTAA CCCAGTGAAA ACATGAAGTC AAGTCTAACT AGTCACTACT
19851 ATTTCACTAC TGCTGACTCC TGATGATCAG CTCCTTTTCT AAGTGCTTAC
19901 TGTCCACTTA TTCCATCATC TGCCTAGAAT TTATGTGAAG GAATCAAAGC
19951 AAAAGGATCA TAAGGCTTCC TTTTTCCAGT ATGTTTTTCC TCCTTTTTGA
```

FIG.3-8

```
20001 AAACTGGGCC AGTTAGCTAT CTCCATTTTT ATTTCATGAA TACATCCCCA
20051 GCGCCTGGTA TATAGTAGAT ATGGAACATT ACACTTTGGA GATATTGCAC
20101 CCATTCTCCA GTTTCTCCAA AGTTACTAAC AATGGTTCCA TCACTGTGCC
20151 AACATATTTT CTTTTTTCAA TATATTGGGA AATAATTCTC CCAGTCTGAA
20201 AATCTGAACA CATTTCATGT GACTTGGTAT CCTCATATGT CTTGGGCTTC
20251 CAATTCTCCA TTCCTAGTTT CAAGTTCATG AACTGTAAAA CAAAGGATTA
20301 GACTAAATCT CTAAAGTTCT ATCCAGATGC CAAATTCTTT TCTCTTTCCA
20351 TGATACCTAA GATAGATGCC AAATATTGTC TTTTACCTGG TGTTTGTGAA
20401 CATGACATCA CATTACAGGA GTAGCAGATA CTAAACTCTC ACTCTGTAAA
20451 ACACTGACTG AGTTCCATGA GCCAGATACT GAAGTGAGCT TGTTCACATA
20501 TGTTCTCATT TAATGCTCAT AACCCTGTGA AGCTGGGAAT TGCTGGGACA
20551 TTTTATTTAT TTATTTATTG AGACGGAGTC TGGCTCTGTC ACCTAGGCTG
20601 GTGTGCAATG GCATGATCTT GGCTCACCGC AACCTCCGCC TCCCGGGTTC
20651 AAGCGATTCT CTTGCCTCAG CCTCCGCAGT AGCTGGGATT ACGGGGCACA
20701 CACCACCACA TCCAGCTAAT TTTGTATTTT TAGCAGAGAT GGAGTTTCTC
20751 CATGTTGGCC AGGTTGGTCA CGAACACTTG ACCTCAAGTG ATCTGCCTGC
20801 CTCAGCCTCC CAAAGTGCTG GGATTACAGG CATGAGCCAC CATGCCTGCC
20851 CGGGACCCTT GTTTTAGAAG GATGACTGCT GCTATAATGT AGAAAGTGAT
20901 TTGGAAGAGG GGAGGAGTGG GGCACGAAAG ATGGTTAGTA GATGGGGGTG
20951 GTAATGCTTA CCTTTCAGTA TTTGGAGGCT TCGGAGTCCT CAAAAATTCT
21001 CTTCCTTGAT TGGAGTCCTC CCAGCCAATA GAGGGCTTCA CACAAACAGT
21051 TTCTTGGGTT TTGAATTGTT TGACCAGAGC TTTCTTCCGA CAAAAGGTTG
21101 GGGTGATTCA TTCACTTACC ACACCTTGCC TGAACATTCA CTTGGGGCTG
21151 CCGGTTATGA AGGCTATTGT TCTCCAGCCT GTCACAGACG CTTTGAAGAC
21201 CTGTGCCTCA GCTGGTTCTA AGGAGTCAGT TTGTTCAGCT CCGTGCCAGG
21251 TTTCCAACTT ATGAAATGTG CTGGAGATTA ACACCTCTCC TGCCATTTTA
21301 TCCCTACTAT AATTGCCAGT CAAAGGATTC CTGCAGTTGC CTCTGGCAGC
21351 CATAACTGAT GAATGTTCTG CCAGCTGCTC TGAGGACCTA GAAGAGCAGT
21401 TTTCTATCCA GGACCAGTTT CCAAGGGTGG GAGGGTGAAA TATATCCTCC
21451 AGTGTGACAT TTCATCTCCC AGTGATGGGT GGCTTGGGCC CTTTGAAGTT
21501 GGCTCTGAGG AACCACACAC TTGGGTCTGA GCAGCCAGCA GCTTATCACA
21551 TCTGGTGATC AATCCTTCAA AGGTTCCTCC TGAAGTCTGA ATTTTTGGAG
21601 GTCAAATGGA TTCCACCTGG GAGGGGCTTC TGCTTCAACT CAGGACATGG
21651 GGAGAAGGCT GTTCCTCTTC CAGGGGGAGG CAGTTTTCAT GGCATTGAGA
21701 TGTCCTCTCA CTTATTCCCC ACCCACCCAC CAAGTCCTTT GTAAGAGGAG
21751 TAGGGGGAGA GGAGAGCGCC TGCAGCCTCC TGCTCACATT CCTAGACACC
21801 GACTCACTGA GCCCGTCGCC GCTGGAACAG CAGAGCTGTG TGAAATGTCA
21851 AGAGGAGTTA TGCTCATAGG CTCCCTGGCC TCAGTCTCTT TGTGGCTTGC
21901 ATATTCTTCC ATTAGTACTG TGTTCATCAC ATGGAAATCA GAGGGTACAA
21951 TTAAAAGATA ATTTGCTAGT CCCAGACTTA ATTTGGGGCC CCCTTCTTGC
22001 CTGATTGAAT TACAGGGGAA CATAATAGAT TTTTGGTGAG AAATAGTTGT
22051 CTGTGTGGCT GGGAGAAAGA TTGCTCCCAG CTCTCCAGCT GGGCAGCCCT
22101 TTCAGTATCC CGTATGTTAT TTCCCCACTT CCAGCCCACC TCACCTCCTC
22151 TGTGGCCCTT GTGTGTCCCC TCGGCTAGGA TCCTGACCTC CTGCTCAAGA
22201 GTTTAAACTC AACTTGAGAC CCAAGGAAAA TAGAGAGCCC TCTGCAACCT
22251 CATAGGGGTG AAAAATGTTG ATGCTGGGAG CTATTTAGAG ACCTAACCAA
22301 GGCCCAGACA GAGAGAGTGA CTTGCTAAAG GCCACATAGC TAGCCCACAG
22351 TAGTTGTAAC AATAGTCTTA ATGATATTAA TGGCTAACAT TTATCAACCT
22401 TTAATGTGTC CCAGACTTTG TGCCAAGGGC TTACATGCAG TGCATTGTCG
22451 CATTCAAACC CAGACAGTCT GGCTCTGGGC CCAGGCTGAG CTTTGGTATA
```

FIG.3-9

```
22501 GCATGGTAGA ACGTTGTCTA TAATGTCTAG TCTGGGTTCA AATCCTGGCT
22551 TCACTTCTCA CATTTACAGC TGAGTGACCT CAGGCAAGTG ATTTAACCTC
22601 CCTGTACCTC AGTTGCTTTA TCTGTAAAGA GAAAAATCAC AGCACTGTGG
22651 AATAGTGGGG GTTAAAATTC ATTCATACAA GTAGTGCTGC AAGCAATGTT
22701 TAATACAGGG TGAGCACCTG TTCAGTGCTT CCTTCTTCTG GCTGCCTCTG
22751 GGGCTAGAGT GTGGTGTCTT CGTGGTATAG ATAGATAGAT ATGGCTGAGC
22801 TCTGCACAAA CACCAAGAGC TGTTCTTCAC TATTAGAGGT AGTAAACAGA
22851 GTGGTTGAGC TCTGTGGTTC TAGAACAGAG GCCGGCAAGC TATGGCCCAT
22901 TGCCTATTTT AATACGGCCT GTGATTGATT GATTTTTTTT TTCTTTTTGA
22951 GACAGAGTTT CACTCTTGTT GCCCAGGCTG GAATGCAATG GCACGAACTC
23001 AGCTCACCGC AACCTCTGCC TCCTGGGTTC AAGCGATTCT CCTGTCTCAG
23051 CCTCTCGAGT AGCTGGGATT ACAGGCATGT GCCACCACGC CTGGCTAATT
23101 TTTGTATTTT TAGTAGAGAC AGGGTTTCTC CATGTTGGTC AGGCTAGTCT
23151 CGAACTTCCA ACCTCAGGTG ATCTGCCCGC CTCAGCCTTC CAAAGTGCTG
23201 GGATTACAGG CGTGAGCCAC CATGACTGGC CTGATTGACT GATTTTTTTA
23251 GTAGAGATAG GGTCTTGGTT TGTTACCCAG GCTGGTCTCA AACTTCTGGC
23301 TTCAAGCAGT CCTCCCTCCT TGGCCTCTCG AATGCTGGGA TTATAGGCAT
23351 GAGCCACTAT GCCTGGCCTA TATGACCTGT GATTTTTAAT GGTTAGGGGA
23401 AAAAAAGCAA AAGAATGCTT TGTGACATGT GGAAATTACA TGAAACTCAA
23451 ATATCAGTGT CCCAGCCTGG GCAACAAAGT GAGACCCTGT CTCTACAAAA
23501 AATAAAAAAA AATAAGCCAG GGCCGGGCGC AGTGGCTCAC ACCTATAATC
23551 TCAGCACTTT GGGAGGCCGA GGCAAGTGGA TCACCTGAGG TCAGGAGTTC
23601 AAGACCAGCC TGACCAATAT GGTGAAACCC TGTCTGTACT AAAAACACAA
23651 AAATTAGCCG AGCATGGTGG CATGCGCCTG TAGTCCCAGC TACTTGGGAG
23701 GCTGAGACAA GAGAATTGCT TGAACCTGGG AGGCGGAGGT TGCAGTGAGC
23751 CAAGATCGCG ACACTACACT GCAGCCTGGG CAACAGAGCG AGACTCCGAC
23801 ACACGCACGC ACGCACACAC ACACACACAC ACACACACAC ACGCTGGGTA
23851 TGGTGGCCAG CACGTGTGGT CCCAGGATGC ACTGGAGGCT TAGGTAGGAG
23901 GATCACTTGA GCTTAGGTGG TTGAGACTAC AATGAACCAT GTTTATACCA
23951 CTGCACTTTA GCCAGGGCAA CAGTGTGAGA CTGAATCTCA AAAGAAAAAA
24001 AAAAAAAAGA AAAAAATCTT TCCATAAGTA AATATCTGTT GGAACATAGC
24051 CATGTCCCTT AGTTTATGTT TTATATATGG CTGCTTTTGC CCTATAATGA
24101 CACAATTGAG TGGCCACGAC AGTCTGTATG GCCTGCAGAG CCTAAGATAT
24151 TTGCTCTCTG GCCCTTTACA GAAAAAGTGC CTTGACCTGT GCTCTAGAGC
24201 CATATGTACC AGGTTTGAAA CTCAGCCTCA CAGCTGGGTG TGATGGCACG
24251 CATCTGTAGT CCCAGCTACT CTGGAGGCTG AGGTGAGAGG ATCACTTGAG
24301 TCCAGAAGGT CGAGGTCAAG ATTGTAGTGA GCCATGATGG CATCACCGCA
24351 CTCCAGCCTG AGTGACAGAG AGAGACCCTG ACTCAAAAAA AAAAAAACAA
24401 AAAAAAAAAA CACCCTCACC ACTTATCAGC TATTTGTCTT GAGAATAGTG
24451 ACATAACCCC TCAGAACCTA TTTCCTAATC TGTTAAATGA GGCTGATGAC
24501 GTTTCCTCCT TTTACTGGCA ATTTAAACAT GATGGATAAT AAATGCTAAG
24551 CACTTAACAC AGGGCCTAGA AGATATTAAC TGCTCAATAA ATGGTAGCTT
24601 CTTAACAGTA TTCAAACCCA TGTGCTCTTA TCACATGCAT TGTTGTCCCT
24651 GTGTCCAGTT GGTGGAATGG GAAAAGGCTC CCTTGTAACC CCATCTACCA
24701 TCTTTATCAG ACTTTCCTGC CATGGTTCAC AGTAAGAGAT AGAAGCTGCA
24751 CGGTGACTTC TGGCTCTTTA CAATGGTGAG CGGTGTGTGC CTGGTAAGGG
24801 AGAGCTGATG TCACTGCCCC AAATCCAGTA GTGAGATCTG AGTGTTCTGG
24851 TTTCCTCCAG CAGCCTTGCT TTTTCCTTTA CAATCCTGCA GGCAGGGAGA
24901 CAAGGGCTTT CTACATGGTA GGCTCTGGTT TGGTCATCGT CACAACTGGG
24951 GGCTGTTCAG GTGGGCTCCC ATTCCAGATA CCTAGGCTTA TCAATCCCTT
```

FIG.3-10

```
25001 TTGGCACCCC AGGCCTTTTT CTCCCTCATG CCCCATTTTT CAGTTTGAAA
25051 AGCATGGTTA TCACAGGACA AGTAGAAGAA GCTCCACTGT CCACTGAGGC
25101 CAATGGATGG TGTTCTGCAT GTGAACACTC AGTGAATAGT GAGTGAATGA
25151 GAGTAACCTG GGCTCCATCC TATTTGCAGA GAGCTTTGGA AAAGATTTTT
25201 CTCCTTAAAG AGCCAGAATG AAGCCTGGTA GTGGGAGAGC TCCAGCTCTA
25251 GAGTCACATG AGCCTACATT TAAATTCCAG CCCTGCCACT GACTCCCTTT
25301 TTGACCTTGA GTGAGTTACC TAATCTCTCT GTACCTCACT TTTCTTGTCT
25351 GTAGAGTGGG AATAATTCCT GTCTCAGAGA AATAAAAGAG TGCATATAGT
25401 GTTTGCCACA TGGAGACACA TCAGGTGTAG GTTAATACTC TGGGCCTTGT
25451 TTCCTTATTT GCAACACAGC CCTGCCCTGG AGTGGAAGTG GCACCTCCCA
25501 TTGGTCAGCT CTTGAGGCTG TCCCCAGGAC AGGCAGAGGG AGGGAATGAA
25551 TGGGAGCCCT AGTGCCAGGA CAGAACAGAT GGCAGCTCAG AGCTAGGATG
25601 GCTCTCTGGA CCTGTCTCTC CTACCAGAGG TCCCCCCGTC TGGTGTGGCT
25651 CTTCCTGGAC CTGGCATCCT CTGCTTTTTT TTTTTTTCCA CCTCCAAGCA
25701 GAATTACTGT CCTGTAGGCA GCTCCTCTGC TTGAGGACAT CTGGGGCCAG
25751 ATATGTTCAC ACTCTATCCT GCCTTGCCCT TCCCTGAGCT CAGGATGGAC
25801 GCTCAATTGG TCCCAGTTAT TGTCTGCAGC GCCTGCCTGC AGCCTCGATC
25851 CAGCCCAGCT CCACCCCTTG CCTGCAAGGT CTGTTTCCTA ACAGCTGCTC
25901 CAACCACACA CCTCGGTTCT GCGGGAGCCC CTCCTCTTCC TCCCTCCCTC
25951 CCTCATTCAG GGGTGGGACT GAAGAAGAAG GCTAACTTGA CAGCAGCGCT
26001 TCTTTCTTAG CTAGTCACCG GCCCCTGCTC AAGAATGCCA GTGTGTGTGT
26051 AGCCTCCACA GAGAGGTCGT TTTCTCGGAG TCCAGAGGGG CCGCCTGAGC
26101 TTCTGAGAAC TAGGGAGGAG CCATCCCAGC CATGAGCCCC TGTGGGAATC
26151 TGCTGGGGGC CAAGTGGCCT GGAGTCCTCA GGCTCCCGCA GCTGCTCCGG
26201 AGGGAGAGGT GAGCTCAGGG CAGCCTGCCT GCAGCCAGAG GTGCCGGGAG
26251 CCCCGGGCCT GTCATGGTGG CCATCTACAG CCGGCCTGAG GCAGTCACAG
26301 ACGGATTTGC AGCTGAGCCT GTCTATCTGG TGTGGGAAGA AGATGGGGAG
26351 TTACTTGTCA GTCCCGGCTT ACTTCACCTC CAGAGACCTG TTTCGGTGAG
26401 TTGGTCTCCG AGTTCCCCTC TCCATCTCTC CTGGCCCCTG GTCCTGAGAG
26451 GAGGGTGGTC TCCCTAAATC TCCTTCTCAC TTAGTCCTTT ACCATCGGTT
26501 CTGCCGGGCA GAAGCCAGCG GAGGTTATAC CCAAGGAGAA TCGGCCTTGT
26551 GAGGTACCCC CATTATGTCC TGGAAGTGGT GAGGGGAGGG ATATACCCAG
26601 AAGGAACTTC TTAGGGAGCT CCAGCTCCCC TTCTATCCCA GACAAACCTG
26651 AAGGAGCCTC CAAAAGATGC CACTGACCTG CCCATTGTAG ATGTTACTGC
26701 TTCCGGGGGG AATAGCCCAA ATAGAGTGCT GTTTCCAGCT CTCACATGTC
26751 TTACCTGCGG GCCATGCTGC CTGCCCAGGA ATTTGTCCCA ACAAGCAGGA
26801 TGGGCAGGTT TTGCCAAACT GTGGAAACTG GCAAGTCCTG GGTGTGGGTA
26851 GCCTGGTACA CAGTAGGCAC CTTATAAACG TTTGTTCTCT TAATGGCAGG
26901 CACATTTGCC TCTGGCCTTG AAGGGCTTCT GAGCTCCCAG GTGAATGTAG
26951 TTGCTGGGGA AAGACCTGGG CGAGTGCTTC TAAGACTGGA GCAATGGGCT
27001 TTAGAGTGTT CCTGAGCTGC TGGGCCAGCC CCCACACCTC CTCAGTCCCT
27051 AGGCCTAAGT ACCTCCACGA GCCTCTCTCT GTGGGGCTTC TCAGAGGGAG
27101 ATGTGGAAAC TCTACCTCTA ACCTGGCTTT CTTTGCTCAT TGCCCCACTC
27151 CACCTCCCAT AGAAACTCCC CAGGGGGTTT CTGGCCCTCT GGGTCCCTTC
27201 TGAATGGAGC CATTCCAGGC TAGGGTGGGG TTTGTTTTCA TTCTTTGGGA
27251 GCAGCCTGTT GTTCCAAAAA GGCTGCCTCC CCCTCACCAG TGGTCCTGGT
27301 CGACTTTTCC CTTCTGGCTT CTCTAAGCTA GGTCCAGTGC CCAGATCTTG
27351 CTGCCGGGAT ACTAGTCAGG TGGCCAGGCC CTGGGCAGAA AAGCAGTGTA
27401 CCATGTGGTT TTGTGGAATG ACCGGACCCT GGTAGATTGC TGGGAAGTGT
27451 CTGGACAGGG GGAAGGGGGA AGGGAACTGG TCCTCAATGC TGACTCTACC
```

FIG.3-11

```
27501 AAGCGCCCTG CTAGACACTT TATCCTTTAA TCTCTCAACA GCCTAAAGAG
27551 ATTATATATC CCCATTTTAC AGATGAGGCA ACCAGTTTCA ACAGAGTTAA
27601 CATATGGAGC CTCACTGGGC AGCTTTTTCT GTCTTCCTGA CTTTCTCTCA
27651 TCCTTCAGGG GGCTGCAGGT TTGTTTTCTT CTCCTAGTGG AGAGGAAATT
27701 CTCAGGTTTG TTTTCCTCTC CTAGCAGAGA GTAAAAAAAG GGATAGTTTG
27751 CCTGACTTGT TGAAGGTGTG GCTGAGATTG TTTTCTAAAG AGCCAATGGA
27801 AATTGATCTT GAGTTTAGGA GAAAGCTTTT ACATGTGGAA TTAAGATGCC
27851 AAGTGTTGAA GTAGCCACAT TTCAGGTCCT CATTAATTTC TCTTAATCCT
27901 GGGAAGGCAG CTTAGGAGAA GGGTTGTTCC TTTAGGAGCC AGGAACTATA
27951 CCCCTTTTAC CCTTGGAGAG GCAGGGAAGC CAGGGAGGAC ACAACTTCTC
28001 AGGAAGAGGA GAAGCTAGAG CAGATAGTGA ACTCTCAACC TGAACCTTTA
28051 AGGGCCAGAC CACTAATGCC ACCCAAGTCC ACCTGCCGTT TGTCTTGTTC
28101 TGTCCCAGGC TTTCTGGAGA ACCTGATCTT CTTGCCCCTA CCCCCAAGCT
28151 CCGTTTGCCC AGCTAGAGTC TGGGGGGTAC TGACTGACTT TCGTAGACAT
28201 TCTTCCCTTC CCCAAATAAG AGGCCACATT CCTGAAGTCA CTTCTGAAGA
28251 GATAGCTGCC ACACAGGGCT CTTTCCCCCC AGGGAGGGAC CACCCAGACC
28301 CTCTGCTCTC CCAGGTATCC GTTACCACAT CACTACCTGG TCAGAAAGCT
28351 GTTTCTGCCA TTAGCCCCTC CCTCTTTTAT TATAGGATAT CCTCAAGGGC
28401 TCCTCTTTGG GCCTCAGTTT CATCCTTGGC AGAAAGTAGA AGCTAGACTT
28451 CTTGGGCTCC TGAACAGGGT CCTTGCTGGA TTCTGTGAAA CAAATTAAGT
28501 TCTTGACCCT AGGCCTCTGG GGGAGTACAA AGTCTATGGG AGTTCTGGGG
28551 CTGTGGTTGC AAGGAAAGTG ACGCAACCAG ATTCCATGGG GACATGATCA
28601 GGCGTGACAT GTGAGGGAGG AAGAGGGAGC AAGGGAATGA AGAATACAAC
28651 TTCTGTGTCC CATACACCCC TGCCTGACAG GCCATACATA CTCAGCAGAG
28701 AATGCACTGT CTTTCCTACC ACACTAGCGT GAGGAGTGAG CTGCAATTAC
28751 CACTGTGCTT CCAAGTAAGA AAATACCTCA AATTGGAATT TACAAAAGAG
28801 GTAAATTAGG GAGTGGCTTT TGTCGGACAT CTTTAAAGCA TTTTTCTTTT
28851 TATAGAATTT CACTTAATGT CCAATACTGA TTTAATGAGC TTGGGTTTAC
28901 ACATTATCTC TTGAAGAAAA CAAATGAACC TTTGTGTTCC AAAGCAATCC
28951 ATGTTTAAAG GGAAAAAATT ATGCATAACT CTGCCCAGCT TCACAGTAAC
29001 CTTTGGCAGG TGCCTTAGGT CCTCTGGGAC TCTTTTCCTT ATCTGAAAAA
29051 TGAAGGACTT GGATCAGGTG AATGGTTCCC AGCTCTGCAA CTTATGTGGC
29101 TCCTCAGAGG CACACAAGCT CTTTTCCATT ATTTGCCAAA TAATGGAGGC
29151 CCTGTCTTTA ACTGCAGTAC AACTACACAA AATACTTGAA ACTACAGTCT
29201 TCCTGGTTTT TGGTTGGAAC TGAATCAGTG CACTCTAGCA ACACTTATTT
29251 CTTGCTGTTC GTAGGCTTCA TTATGTGTTT GGTTAATTTT TTAAAACAAC
29301 AATAACATAT TCCATAATAA TTACAGCTTA ATTGGCAGAC TGTTTCAGTC
29351 TATAGGATCT GCAGGAAGGA GGAGTAATAA AGGGATTTTT GACTGAGCTC
29401 TTATGGAACA GAGTCTCTCT AGGCCCCTGT CATATCTGCC CTTCTGGGCC
29451 CTGGGGAAAA GTTGGCATCC CCAGTTGTGG TGCTCTCCAG GTGCCCTCAG
29501 GCTGTGGTGG AGGGAGCTTC CCATTCTCTC CTTCAGCCCA CTCAATTCAG
29551 AGGCTAGGGG CTGAAAGAAG CTTCTCTACA ACTGGCTGTT CACTGGGAGG
29601 TTAAGGGATG ACCATCCAGC CAGGCCTTCC TCAGGACATG GGAGGGCTTA
29651 TGCTTTAACA TGTGTAAATC CACTGCAATA ATGACTGGTT CTTTTACCCC
29701 ATAAGGTTGA GAATTTACCT GTAAACATTT TTGTCTGAAG AATTTGGATG
29751 TAAGTGAGGG CTGGGCCTCT ATCTTATCTC ACTTGGCTTC TCTCAGCACA
29801 GCACCTTGCC TGCTTGTTCT TACACATCCT AGATGCACAG TAACTATTTC
29851 CTAATTATTA GAAATCTATT AGAATCAATT GATTTCAGCT GGGCTTGGTG
29901 GCTCCTTCCT GTAATCCCAG CACTTTGGGA GGCTAAGGCT GGAGGATCAC
29951 CTGAGTCCAG GAGTTTAAGA CCAGCCTGGG CAACATAGGG AGACCCTGTC
```

FIG.3-12

```
30001 TCTACAAAAA ATAAAAAATT AGCCAGGCAT GGTGGTGTGC ACCTGTAGTC
30051 CCAGCTACTC AGGAGGCTGA GGCAGGAGGA TCTCTTGAGC CTGGGAGGTC
30101 AGACTACAGT GAGCAATGAT TGTGCCACTG CACTCCAGCC TGGGTGACAG
30151 AGTAAGACTC TGTCTCTTAA AAAAAAAAAA AAAAAAGTTG ATTTCTATTT
30201 GGATAGATAA ATAATTCATT TTAGGACCTT TCTTTTTCAC TTACAGAAAT
30251 CTGTTTCATT CTGGGCTGAG AAGCAGGTCC ATATTGCTAG GCATAGGAGA
30301 AAAAGGGGTC TGTCTGCATT TGCCCTTGGT GGTCTCAAAT TGGGGAGGGA
30351 AAGAAATGAA CACTTACTGG CTACCTTCTG TGAGCCAGGC ATCATGCAAG
30401 ACATCTGTAC ATAATTTAAT TCTCATAACC CCATAAGATA TTATTAGCAA
30451 TGTACAAGTG AGGAAACTGA GGCTCAGAGT CATGAAGTAA CTGGCCTTGG
30501 GTGACACAGA TGGTAAATGG CAGAGAAGGA ATATGGATCC AGGTCTTGAA
30551 AGAGAAAATC TCAACTGATT ATCTTTTTTA AAAAACTCAT ATGTTCTCTG
30601 CTGACTCAAA AGGTCTCTGT GTGGATCTGG GTTGACCCAC TGAACTGACC
30651 ATCAGGGTTC CATGCACTTT GTATCTGCCC AAGCCCTCAG AACCCCTCAG
30701 TAATGTTTTG GAAGATGAGT TTTGGAGGTT GTCCTTAGGC ATAGCCTCAG
30751 CGTATGTAGG CCTCTAGGTG ATCTCCCCTA ACCTGAGGAT TTCAGCTCAA
30801 TTCACTCTGG CTCCTCAGGA CAGTGGGATG ACTGGTTCAG ACCTCAGCTT
30851 TACCACCTCC CAGCTGGGTA CTCTTCTACC TACAGCCAGG GCAGATTTTG
30901 ACTTTCACTT GAAACTTCCA AAAATTGAAA GGTAGAAAAA CAGCCTTGGC
30951 TTTGGGAAGA ACGTATGATG TCCATGGCCT CTAAGCATCT GAGGTGGGAC
31001 ATGTTCGAGT AGCACCTTAC AGTTCCAAAG TGTGTTCTGG GTTCTTTGTT
31051 TAAAAGAACA GAGACTGCTG GGGAATTGAA CACTGTGAAG TATATGAAGG
31101 AGGAGAATTG TGCTATTTAA CATTCAGTAC TTGGGCTAAA GGAGAAGCAT
31151 CACGAAGTGT TAACACTCAA AGGGTCTTGA GCTGTCAGGG CTCCAGCTTC
31201 CTTATTTTCA CAGGTGAGAA TCCTGAGGCT CAGCTGTTGA GATGTGCTGT
31251 CTCACTCCGG TGACATAGTA CAGTGGATGT GGCTTTGCAG CCAAGCACAC
31301 ATAGCTTCAC ATTCCAGCTC CATCAATTAT GTATTGGGCA GCTTTGCAGA
31351 ATGATTTGAC TTTAACTCTG CTTTTCAGTC TTCTGTAAAA CAGGGATAAT
31401 CCTGCTACCG TAGGGTTGTC AGGATTAGAG ATAATATAAA TAAGGTACCT
31451 CATATAGGAC CTGGATTATG GCTGGCATTC AATAAATAGT AGCTGTTAAT
31501 TGATAGCTAA GCTAGAACTC TGAAGTCTAC CATGGCAACT TCTTAAGTGG
31551 TCTGAGAACC CAGTTGTGTT CTGTGGCAAA ACACAGCTTA GGGATCCATA
31601 CCCAGCCCTC CTGTCAGCTG TTCACCTTCC AGTTCTTCAG AGACATGTGT
31651 GGCAGTGACT TTGGCCACAT AGCTGGCTGT GCCCTTTAAA GGCATTCCTT
31701 GACACAGATA TGTGGACTGG TGACGTTGCT CTCCAGCCAG GTGTTCTTCC
31751 CAGCAGGCTG GCCTGGCTGT CTCCTGCATG CCTGTACTTG TTTGTCTCCC
31801 TGCTCCCTCT CCTGGGCCTG GCCAGAGCTA CTTGCAGCAA ACAAAAGCAG
31851 GATATTGGCA ATGGAAAGGA GGGTGTGTTC TGGTGCTCCC ATGCCCTGCG
31901 GCGCACATAC CATTGCAAGG GCGTAACAGA GCCCAGGCCT GCATTTGGGT
31951 GCAAATAAGT CTGCACACAG AAGAAAAGAA GGACCTGGTG ACCAGGAGCC
32001 ATGGAACCCT TGTGCTCCCC TACCTGGGCT ACTGGTTCTT GCCACTCCTA
32051 CCATTTTCAG TTTGGAAATA TTTGTTAAGG CTTTGCTCTT CCAGGTCCTT
32101 TGCTTGGTGC TGAGTCTACC AAGAGTAAGT GGGATGCTGT TTTTGTCCTC
32151 AGGGAGCTAA CAGTCTAGTG AAGAAGAAAG ATGGTTGCCC AGGAACTTCT
32201 AAGTCAGAAG GCAGGAGGCA AGAAGGAAGC CCCTGCTCCT ACTGCCAGCC
32251 CTCTGTTGGG CACCCCATAG TTCTTCAGAA CCACATTTAA TCCTCACTGC
32301 AGGCCAGGCA TAGTGGCTCA CACCTGTAAT CGCAGCACTT CGGGAGGCCA
32351 AGGCGGGCAG ATCACTTGAG GTCGGGAGTT CGAGACCAGC CTCACCAACA
32401 TGGGGAAACC CCGTCTCTAC TAAAAATAGA AAAATTAGCC GGGTGTGGTG
32451 GCATGCGCCA GTAATCCCAG CTACTCAGGA GGCTGAGGTG GGAAAATCAC
```

FIG.3-13

```
32501 TTGAACTCGG GAAGCAGAGG TTGCAGTGAG CCGAGATTGT GCCACTGCAC
32551 TCCAGCCTGG GCGATAAGAG CAAAATTCCA TCTCAAAAAA AAAAAGAAAA
32601 AAGAAAAAAT CCTCACTGCT ACCTTGAAAG TAGGTGATGA CATTGCCATT
32651 TCACAAATGA GAAGTGAAGG GGCTAGCCCA AGATCACTTA GGTGGTAAAT
32701 GGTGGTGCTA AGATTAGAAC CTCAGATCAT CTAGGGAAAA ACACAGATAT
32751 GCACAGAGTT AAGGGGACCC AGGGTATTGT TTGTCCTCTT GTTTCACAGG
32801 TGGGGAAACA ACCCAGAGAG GGAAAGGGGC TTGTCCAAGG CAATTTAGCA
32851 CCCAAGAACT TGAACCCATA TCTCTCTCCT CCTCATTTAG AGCTCATCCC
32901 ACATGTATCT TATATTGAGA GGAGTGTGAG CCACATACCA AGAACAGTCT
32951 TCCCCTCTGC CTCCAACCTC ACTGTGCAGT TTTGAGACAC TTCACAGCCA
33001 TACTCTTCAT GCCATACCCA GCCCTTAAGA CCCTGAAGTT CCCCTTCCAT
33051 AAGACAAGTA GGAAAAGCTA TAGGGTAAAA ATAGCCATCA GTGTTTGTTG
33101 AGCACCCAGG AGGAATTGGG CACTCCAGAA AGATAAAGGG ATTCTCAGGG
33151 ACTTGCTTCT CTAGACTTCC CTAGCTCAGC TGCTTCAACT CATTCCTGCC
33201 CCTCTTCTCT ACCTCCCGCA GTGCTCAGAA GTAGTAGAAC TCACTGTGGC
33251 CTCTCACCTT GCATTGTTGA GTTTTATTTA GACTTTCTCT TCCTCAACTC
33301 TTCATAAGCT CATGAAAGGT GAAGTAGGGT GCCCTGTGTA TTTATCTTTT
33351 ATATCTGCAG TGCTTAGCAA GTTATAATAA TGCACTTGCC TGGCAAAAGG
33401 CTTTCTCTCA TACATTAGCT TATTTCCTCT TCACATTGGC TCTTTGTAGT
33451 AATAGGATGC TATTAGTTAT TTTCAATGAG AGAAAGCTAC TAAGAGAAGT
33501 TGTCCAGCTA GTGACAGTAA GTGGCTGATA AAGTGAGCTG CCATTACATT
33551 GTCATCATCT TTAATAGAAG TTAACACATA CTGAGTTTCT ACTATATTGG
33601 GTCTTTTTTT TTTTTTTTTT TTTTTTTTTA GAGACGGAAT CTTGCTCTGT
33651 TGTCCAGGCT GGAACGCAGT GGTGCAATTT TGGGTCACCA CAACCTCCGC
33701 TTCCCAGGTT CAAGCGATTC TCCTGCCTCA GCCTCCTGAG TAGCTGGGAC
33751 TACCAGTGCA CGCCACCACG CCCGGCTAAT TTTTGTATTT TTAGTAGAGA
33801 CAGGGTTTCA CCATGTTGGC CAGGCTGGTC TTGAACTCCT GACCTTGTGA
33851 TCTGCCCGCC TCAGCCTCCC AAAGTGCTGG GATTACAGGT GTGAGCCACC
33901 GCGCCCTGCC TATATTAGGA CTTTTATATA AGCTATCTCT AGCTAGCTAG
33951 CTAGCTAGCT ATAATGTTTT TTGAGACAGA GTCTGACTCT GTCACCCAGG
34001 CTGGAGTGCA GTGGCGTGAT CTCGACTCAC TGCAACCTCC ACCTCCTGGG
34051 TTCCAGTGAT TCTCCTGCCT CAGCCTCCCG AGTAGCTGGG ATTATAGGTG
34101 CATGCCACCA CGCCCAGCTA ATTTTTTGTA TTTTTAGTAG ACCAGGTTTC
34151 ACCATGTTGG CCAGGCTGGT CTCGAACTCC TGACTTCAAG TGATCCACCC
34201 GCCTCGGCCT CCCAAAGTGC TGGGATTATA AGCATAAGCC ACTGTGCCCA
34251 GCTGCTCTCT ATATTTTTAA TACATATTAT TTCCATTAAT TTTCACAGCA
34301 GTTCATTTTA TAGATGAGGA AACTAGGCCA GAGAAGTAAA ATATCTTGCC
34351 CAAGATGATG TAACTAGTAA GTGGCAGGAT CAAGATTCAA ACCAAGCAAT
34401 GTTCAAACCT CTTGGAAGCA AGAATGTGGC CACTGTGGAA GGTGCAAGGC
34451 CTTGACAACA AGAATAGGGA AAAGAAGGAA CTAGAAGGAA AGAGATGGCA
34501 TGGGCTCAGC AGGCCAGGGA GCTCTTAGCT GTGTGTGTTG GGAAGCTCAG
34551 AAGGGAGGAA GAGGTTGTCT GTGCAGGTAA GTCCTGAGAA CACACCAGAC
34601 TTTTGAGAGG TGGAGCTTCA TAGCCAGGTC ATTAGGGGAG AAGGGAGCTA
34651 TAGATTTTTT TTTTTTTTTT TTTTTTTTTT TTTTTTTTAG AGACGGGGTC
34701 TTACTATGTT GCCCAGGCTG GTCTTGAACT CCTGGGCTCA AGTGATCCTC
34751 CCACCTCAGC CTCCCAAAGT GCTGGGATTA GAGGCATCAG CCACCCCGCC
34801 CAGCGAGCTA TGGATCTAAC ATGTACATCT TACACAGTGC TAATAGAATG
34851 TTGGGTTTCT TCCCCAATAT TTTATTTTGA AAAAAAATTC AAATATATAG
34901 AAAAGTTGAA AAATGTAGTT CAAAGAACAC CTACATACCT TTCACATAGA
34951 TTCATGATTT GTTAATGTTA TGCCACTTTG TATATATCTC TCTCCCTCCT
```

FIG.3-14

```
35001 ATCTGTATAC TTTTATTTAT TTATTTTTGC TGAACTATTT CAGAGTAACT
35051 TAAAGGCATC TTGATTTTAC CCTTGAACAG TTCAATATGT TTCTGCTAAG
35101 AATTCTCCTA TATAAGTCAG ATATCATTAC ATCTAAGAAA ATTCACGGCA
35151 ATTTTACAAT ATAATATTAT AGTCCAAATC CATATTTCCT CAGTTGTTCC
35201 AAAAAATGTT CATGGCTGTT TCCTTTTTTA ATCTAAATTT GAATCCAAGT
35251 TTGAGGCATT GTATTTGGTT GCTGTGTCTC TAGGGTTTTT AAAATCTGTG
35301 CCTTTTCTTC TCCCCATGAC TTTTTAGAAG AGTCAAGACC GGTTATTCTT
35351 ATAGAATAAC CCACATTCTA GATTTGCCTG ATTAGTTTTT TTATACTTAA
35401 CGTATTTTTG GCAAGAACAT TACATTGGTA ACGCTGTTGG TGATGGGTCA
35451 GTTTTGAAGA GTGGAGATGA TTAAACTGCT TTTGTTCATT GAAGTATCTG
35501 TCAAGACCAG AGATCCTTAA CTGGTGCCAT AAATAGGTTT CAGAGAATCC
35551 TTTATATATA CACCCTGTCC CCCACCTAAA TTATATACAC ATCTTCTTTA
35601 TATATTCATT TTTCTAGGGG AGGCTTCTTG GCTTTTATCA AATTCTCAGA
35651 GGGCCCCAAG ACCCAAAGAG GTTATGAAAC ACTAGTCTGT CCACTGAGGC
35701 AGGCAACACA GAGCTGGTTT CTGGGGCCTT GTTCAGTCTG AACCAGCTTC
35751 CCTTGGGGAG ATAGCACAAG GCTGTAACTT TGCCCCATCT TGGCTTTGGA
35801 TCAAAGAGGA CTGTCCATTT TGTTGTCATA CCTAGGAACC AGGGACAGCT
35851 TATGTGGCCT GGTTCCAGGG ATCCAGGAGA ATTTCAGTTC TTGTCTTGCC
35901 TTTCAGGTGT TCAGAATGCC AGGATTCCCT CACCAACTGG TACTATGAGA
35951 AGGATGGGAA GCTCTACTGC CCCAAGGACT ACTGGGGGAA GTTTGGGGAG
36001 TTCTGTCATG GGTGCTCCCT GCTGATGACA GGGCCTTTTA TGGTGAGTGA
36051 ATCCCTTCAT ATCTGCCCCT CTTGGTCTTC AGAGTCCATT GACAGTGCTT
36101 CCAGTTCCCT GTGGCCTGTT AATCTTTTAG TCTTTCCATC AGCCAGGGCA
36151 TCTCCCTTTA TTTATTCATT CATTCAACTA GCAGGTATCA ATTGAGCACC
36201 TACTAAGTGA AAGGTAAGAT CCTTCCCTCA AAGACTTAAT AGTTGAACGT
36251 TGGGAGTGGG AGGAGAGGCA GGCAGAGAGG AGACACAATA TAGTTGGATA
36301 AGGACCTCCA AGGAGAGTGT TACAGGCTGA GAGGAGGATA TACTTAGGTT
36351 GTCTTTAGGG AATCAGAAAA GGAGACTCTG GAATAGGCTG GCAGAGAGAG
36401 GGGCTACCTC CTATACCTGC TCTGGACAAA CGACTTTAAG CATAGTGACA
36451 GATTTGCCAA CCCTGTATTG GAAGAACTGA TCTTTTTTAG TGGGGATGAT
36501 TACTTCTGGG GATTTCTTCT CATAACTGAG ACCAAAACAG TTTTGTGCAG
36551 TCTCAGAAAT GACAGGAGGT ACCAATCTGA CACTTCCTTT GGAAGCTCTA
36601 GGGCAGAGAG TGAAAGAGTG GATTTTGACG GGGGCCTTGC TTGGAGGTCA
36651 TTCACCCACC CCTGTCCTCA CTCCAGCAAC AGTGATAACT CACTTCCTTC
36701 CTCCCTTTGT ACACCCTTCT CCCCACCTGC TCACAGGTGG CTGGGGAGTT
36751 CAAGTACCAC CCAGAGTGCT TTGCCTGTAT GAGCTGCAAG GTGATCATTG
36801 AGGATGGGGA TGCATATGCA CTGGTGCAGC ATGCCACCCT CTACTGGTAA
36851 GATAGTGGTC CTTTGTCTAT CCTCTCCCAT ATAAGAGTGG CTGGCGGGGA
36901 GGGACAGTGG CAGGGTGAGT TGGGCAGAAG GAGTGTTAGG GTAGTCAGAG
36951 CATTGGATTC TTACCACAGC AGTGCTCTTA ACCAGCTCTT TAACTTGTAA
37001 GCAGAATGAT TTACACATGT CTCTACCCTT TTTCCTTACC AACCTTGAAA
37051 ATGTCTTCAC TCTGCCCTGC AATCCTCCCA GTGGGAGGCA CTCTTCAAGG
37101 ACGATCCCAG AACATTAAAG TCAAAGACCC CTTAGAGCTC ACCCTGTCCA
37151 ACCACCTTGG TTGATAAAAG AAGTCAGCCT GGGGCCCATG GAATAGAATA
37201 GTACAAGGGC AAGGTTCTCA TTGTGAGTCA AAGGTAGAGT GAAGAGAACC
37251 CAGACCATCT CACCCCAACC CAGGCCAGTG TTTTTCCAAA TATACCACTT
37301 GCTGCAGATC TAGCTCAGCA CCCCCAGTCC CAGCCCACCC TGAGAACCCA
37351 GGCTCCTCAT TCTGAGCAGC CAGCTAGAAT CATGACAAAG AGGGTGGTAG
37401 TGAGACTATG GGTACTGTTG CTTAAAGCCA CATGGTGCAG TGGTTGCTGG
37451 GGGGCTTCTG TGTGGGACTC TAGCATCTTA TTCCCCCCTG TGCCCTCTCC
```

FIG.3-15

```
37501 CCAGTGGGAA GTGCCACAAT GAGGTGGTGC TGGCACCCAT GTTTGAGAGA
37551 CTCTCCACAG AGTCTGTTCA GGAGCAGCTG CCCTACTCTG TCACGCTCAT
37601 CTCCATGCCG GCCACCACTG AAGGCAGGCG GGGCTTCTCC GTGTCCGTGG
37651 AGAGTGCCTG CTCCAACTAC GCCACCACTG TGCAAGTGAA AGAGTAAGTA
37701 TTTTGAGAAC CCTTCAGCAG GGGTTCTTGA GCAGAGTCTG TAAATGGGCC
37751 TCAGAGGGCT TAGACCTCCA AAGTCTCATG CAGAACTCCC TTTATTCTCA
37801 TCTCATATCT TTCTCCTGGA CCCCACTATG CTGTAACCGT ACCTGGGCCT
37851 TGGCACTTAC TGTTCTCTCT GCCCAGGCTA CTTCCTACCC GATACTTAAG
37901 GCAAGAATCA CTCACCTTTC AGGTGTCAGG TTTCAGGTCA TGTTTGCTCT
37951 TTGAAATCAT CTGGCTTGAT TATGTGTATT AGTTGTTTAT CTTCTATCCC
38001 CTCCACTAGA ATGTAAATTC CAGAAGAAAC TTGCTGTCTT ATTCAGTGCT
38051 GCATGCCCAG GGCTTGGAAG AGTACCTGGC ATATAGTAGG AGTTGATTGA
38101 TTATTATTTT GTCAGTCGAG AGAATGAATG GAGAAAATGT GGTCCATGGC
38151 CCAAAAGAAG TTAAGACCCT ATCCTAGATT CAGGCCAGAG ACCAGATGGA
38201 GAAAGAGTCT GTGTCTATCT AATACCAGTA ATGTCGTACC TCTGGCCGCT
38251 TACCATGTAA ATATTGATTG TGTATCTACC ATGTGTTGGA CACTAGGCTA
38301 GTGCTTGCAC AGCAGGTGAA AGATACTAGA GTTTGGGAAG TCAGGAGGAG
38351 CTAAGGTCTG TTCTACAACC TTATTAGATG AAGAGGAGAG GGAATTGTGT
38401 TCAGGGCAGA GGGAGAAGCA TTTCTCCAAA AGTAGGAGTC TTAATCATGT
38451 CTGATGTAGG TTGAGTGTGG CCAGAAAAGG GGCTGTTAAG TATAGAGGGC
38501 CTGGATTATG AAAATCCAGC AGATCCATTG AGAGTTTAAG CAGCAAGGTG
38551 TTGTGACCAA GTTAACATTT TAGAAGGATC ACTGGTATGG AGGTTGGATT
38601 GGAGAGGGGA AAGCCTAAAG GTATAGAGAC TAGTTAGGAA GCTATTGTAG
38651 GCTGGGCATG GTGGTTCATG CCTGTAATCT CAGCACTTTG GGAGGCTGAG
38701 GTGGGAGGAT TGCTTGAGGC CAGGAGTTGA AGACCAACCT GGCCAACATA
38751 GCAAGACCCC GTCTCTGTTT TTCTTAATTA AAAGAAAAGT CCAGACGTAG
38801 ACATAGTGGC TCACGCCTGT AATGCCAGCA CTTTGGGAGG CCAAGGTGGG
38851 CAGATTGCTT GAGGTCAAGA GTTTGGGATT AGGCCAGGCG CAGTGGCTCA
38901 CGCCTGTAAT CCCAGCACTT TGGGAGGCCG AGGTGGGCGG ATCACAAGGT
38951 CAGGAGATCA AGACCATCCT GGCTAACACA ATGAAACCCC GTCTCTACTA
39001 AAAGTACAAA AATTAGCCGG GCATGGTGGC GGACGCCTGT AGTCCCAGCT
39051 ACTCGGGAGG CTGAGGCAGG AGAATGGCGT GAACCTAGGA GGCGGAGCTT
39101 GCTGTGAGCA GAGATCACGC CACTGCACTC CAGCCTGAGC GACAGAGCGA
39151 GACTCCATCT CAAAAAAAAA AAAGAGTTTG GGATTAGCCT GGCCAACATG
39201 GCAAAACCCC ATCTCTACAA AAAGTACAAA AAAATTAGCT GGGTATGGTG
39251 GTGCGCGCCT GTAATCCCAG TTACTCAGGA GGCTGAGGCA TGAGAATTGC
39301 TTGAGCCTGG GAGGTGGAGG TTGCAGTGAG CCCAGATCAT GCCACTGCAC
39351 TCCAGCCTGG ATGACAGAGT AAGATGCCAT CTCAAATAAA AATTAAAAAC
39401 AAAGTTTAAA AAAAAAATAG AAGCTATTAC CGTGATCCAG GTAAGAGATG
39451 TGAATAACTA CAATGATGGA AAGAAGGCAG AGTTCTTAGA GATGGGAGTA
39501 GGAGAGATGA GGGAACTCCA GATTGGGAAG ATGATGTTCA AGTTTCTGGC
39551 TTAGGCCACA GGGTGAGTGG CAATTCCCTT CACTGAGATG GGGCATCCTG
39601 GAAAAGGTGT TGCCTTTCTG TGTGGGTATC CTGGGCCCCT TAGGGGCCAC
39651 TGGTGGCCTG GGACCTGGTA AACCTTCCCT GCACAAGCAG AATTGGTCAA
39701 GCAGGTTTTT AGGACATCTT TACCCTGCCT CAACTCTTGT CTGGCCCAGG
39751 GTCAACCGGA TGCACATCAG TCCCAACAAT CGAAACGCCA TCCACCCTGG
39801 GGACCGCATC CTGGAGATCA ATGGGACCCC CGTCCGCACA CTTCGAGTGG
39851 AGGAGGTAGA GTGTGTGTCT AATCTGTCTT GTGAGGGTGG GACATGGAAC
39901 AGATCCTCTG GGAAATCAGG CTGTAGCCTT TACCTTTTCC TACCCCCAGC
39951 CCATCTCTTT GTCTTAGCAT TGAGCCTGTG ACCACTGGTG ACCTATTTCA
```

FIG.3-16

```
40001 GCGTAACAGG TTCCCAGGGT AGCAGGGATG GTTGATGGAC GGGAGAGCTG
40051 ACAGGATGCC AGGCAGAGGG CACTGTGAGG CCACTGGCAG CTAAAGGCCA
40101 CCATTAGACA AGTTGAGCAC TGGCCACACT GTGCCTGAGT CATCTGGGTT
40151 GGCCATGGGT GGCCTGGGAT GGGGCAGCCT GTGGGAGCTT TATACTGCTC
40201 TTGGCCACAG GTGGAGGATG CAATTAGCCA GACGAGCCAG ACACTTCAGC
40251 TGTTGATTGA ACATGACCCC GTCTCCCAAC GCCTGGACCA GCTGCGGCTG
40301 GAGGCCCGGC TCGCTCCTCA CATGCAGAAT GCCGGACACC CCCACGCCCT
40351 CAGCACCCTG GACACCAAGG AGAATCTGGA GGGGACACTG AGGAGACGTT
40401 CCCTAAGGTG CCACCTCCCA CCCTGGCTCT GTTCTGTCCT ATGTCTGTCT
40451 CTCGGATGAA GCTGAGCTGG CTTTCAGAAG CCTGCAGAGT TAGGAAAGGA
40501 ACCAGCTGGC CAGGGACAGA CTATGAGGAT TGTGCTGACC CAGCTGCCCC
40551 TGTGGGGATC ACAGTTTACA GCCAGAGCCT GTGCGGACCC AGCTGTCTGC
40601 CAGGTTTCCT TAGAAACCTG AGAGTCAGTC TCTGTCCACT GAACTCCTAA
40651 GCTGGACAGG AGGCAGTGAT GCTAAACCCT GAAGGGCAAC ATGGCCTATG
40701 GAGAAAGCAT GGAGCTCAGA GCCTGGAGTA CGGGCACAGA TAGGATTGAA
40751 TAAATTGTGT AGAAAGACTT TGAAAACAAT AAAGCAAAAG ATGAATGAAC
40801 GTTTTTTTTA GACTTGAGGG ACCAACAACC CCCAAACCCC AGATTCTGCC
40851 AGGTCCATGG GGAAGGAGAA GTTGCCTTGA GTGGAAGCCC CAAGTAGGGA
40901 GACTTACAGA AAAGAAGTCA AGAGCACTGG CTCCCAGGCA GAAATACTGA
40951 TACCCTACTG GGGCTTCAGG CTGAGCTCCT CCCTTCACAA ATCACTTCAT
41001 CTCTCTGAGC CTGTTTCTGC ATCTGTGACA TAAGATGGTA AGATAAAGGT
41051 GGCTGTCTCA CCAATTATGT AAGGATTAAA TGTGGAAAAG GACATAAAGT
41101 TGTATAGTGC TGCCATAGGG ACAGTGTTCA GTAAACGTGA CACATTCTTA
41151 GTATCACTAA GAATCAGGTT CTTGGCCAGG CACCGTGGCT CATGCCTGTA
41201 ATCCCAACAC TCTGGGAGGC CTAGGTCGGA GGATGGCTTG AACACAGGAG
41251 TTTGAGACCA GCCTGAGCAA CATAGTGAGA CACTGTCTCT ACAAAAAAAA
41301 AATAATAATA ATAATTGTTT TTAATTAGAT GGGCAGGGCA CTGTGGCTCA
41351 CACCTGTAAT CCCAGCACTT TGGGAGGCCA AGGCCGGAGG ATTGCTTGAG
41401 GCCAGGAGTT CAGGAGCAGC CTGGGCCACA TTCCTGTCTC TACAAAGAAT
41451 AAAAAAGTTA ACTGGGCATG GTGGCACATG CCTGTAATCC CAGCTACTCA
41501 AGAGGCTGAG GAGGAGGATT GCCTGAGCCC AGGAGTTCAA GACTGCAGTG
41551 AGCCTTGATC ACACCACTGT ACTACAGCTT GGGCAACAGA GTGAGACCTT
41601 GTCTCCAAAA AAAAAAGTTT GTTTTTTTTT ATCCACTCTC CTCACCAAAC
41651 AAACTGAGTA AGTTAGAGCC CTCTCAGCTG GCATGTGTTG GAAACAGTGC
41701 CCTCTCATTA AAGTGCTGCC CTCACTCCCA TTGCCTCTTG GCCTTGGTCA
41751 GTATGATGAA ATTAGTGGGA GGCAGGGCAA CAGAGGGCAG GGAAGAGCTA
41801 GAAATCCATG GCCTGGAAAA GGGAAGATTT GGGAGTGGCC AGGTATCTGT
41851 AGAGCCACCA TGCAGAGGAG GGGGGCAGCT AGCCTTGTGT GCTCTGGTGG
41901 GCATGGTCAG CAGGAGGCAG AGCAAAAGGA CAAGGGTAAG TAAACCTGTA
41951 GGTCGGGACA AGCCAAGAGC CATCCAGCGT CAGTCCTCTC TGGGTAGCCC
42001 AAGTAAAGCA GGAGCATACC CCAGAGAGAA AGTTCGCAGG GCTGTTCACC
42051 TGCAGTGCTG TGGACTTCAA CCTTCTTGTT CCTTCTTCAG TAAGTGAAAA
42101 TAACAGTCAT TGACCATGAC TATTATCGAC CGCTTTTGAA AATGTAAACA
42151 TAGTGACTTT ATTGCTGTAA AAATCATACG TGTTTATCAT CTTAAAATTC
42201 AGGAAACATG GACAGGTACA AAGATGTGCA AAATATCATC CAAAATCCCA
42251 TTTGCTGGCC AGGCACGGTG GCTCACGCCT GTAATCCCAG CACATTGGGA
42301 GGCCGAGGCG GGCAAATCAC TTGAGGTCAG GAGTTTGAGA CCAGCCTGGC
42351 CAACATGGTG AAACCCTATC TCTACTAAAA ATACAATAAT TAGGCTGGGC
42401 GCAGTGGCTC ACGCCTATAA TCCCAGCACT TTGGGAGGCC GAGGTGGGCG
42451 AATCACAAGG TCAGGAGTTT GAGACTAGCC TGGCCAATAT GGTGAAACCC
```

FIG.3-17

```
42501 CATCTCTACT AAAAATACAA AAATTAGGGC CGGGTGTGGT GGCTCACGCC
42551 TGTAATCCCA GCACTTAGGG AGGCCGAGAC AGATGGATCG CGAGATCAGG
42601 AGTTCGAGAC CAACCTAGCC AACATGGTGA AACCCCATCT CTACTAAAAA
42651 AATACAAAAA TTATTCGGTT GTGGTGGCAC ACGCCTGTAA TCCCAGCTAC
42701 TTGGGAGGCT GAGGCAGGAG AATCTCTTGA ACCTGGGAGG CAGAGGTTGC
42751 AGTGAGTGGA GATCCCGCCG TTGCACTCCA GCCTGGGCGA CAGAGTGAGA
42801 CTCCATCAAA AAAAAAAAAA AAAAAAAAAA AAATTAGCCG GGCGTGGTGG
42851 CGTGCACCTA TACTCCCAGC TACTTGGGAG GCTGAGGCAG GAGAATCGCT
42901 TGAACCTGGA AGGCGGAGGT CGCAGTGAGC CGAGATCGTG CCATTGCACT
42951 TCAGCCTGGG CGACAGAGCG AGACTCTGTC TCAAAAATAA TAATAATAAC
43001 AATAACTAGC CGGGCCTGGT GGCACATGCC TGTAGTCCCA GTTACTCAGG
43051 AGGCGGAGGC ATGAGACTCA GGTGAACTAG GGAGACAGAG GTTGCAGTGA
43101 GCCAAGATCA CACCACTGCA CTCCAGCCTG GTTGACAGAG CGAGACTCTG
43151 TCTCAAAAAA AAAAAAATCC CATTTGCTCA TTTTTTGGAT ACTAGTATAA
43201 CTATCACTCT AAACCAGTTA GTACTTAAAT CAAGCAGATA TGGGAGATGG
43251 TGAATTACCA TCTACAGTGT TGTCATATAT GTCACATACT GAGCATTATC
43301 AGCTAGTAGA ATCTAGTTAA TTGTTCTATG TGTGATGTAT GCAGAGTTCC
43351 CATTTTGAAT GTGTTTTTAC TATGCTTAAA TAAATGACTG ATGTCAGCAA
43401 CCCCAAAATG ATACATCTGA TGTAAGAGCC CCTGTTCCCC AATAATAACA
43451 TCTAAACTAT AGACATTGGA ATGAACAGGT GCGCCTAAGT TTCCTCCCTC
43501 CAGGGTTTCT TGGCCGGTCT CTGAGGACTA CACATCCCTA CTCCCGTCTT
43551 TCCTCATCTT CAGGCGCAGT AACAGTATCT CCAAGTCCCC TGGCCCCAGC
43601 TCCCCAAAGG AGCCCCTGCT GTTCAGCCGT GACATCAGCC GCTCAGAATC
43651 CCTTCGTTGT TCCAGCAGCT ATTCACAGCA GATCTTCCGG CCCTGTGACC
43701 TAATCCATGG GGAGGTCCTG GGGAAGGGCT TCTTTGGGCA GGCTATCAAG
43751 GTGAGCGCAG GCAACAATTG CTTTGCTCTT CTGCCCCCAG TCCCTCTGTC
43801 ACTGTCTTTC GGGGATTTCT CATCACTTGG CCCCACCCCA CACCATGCAG
43851 GATGCCAGGC CTCCTTCCTG GCTTTGGGTG TTGGTGTGAG AGGTATCCTT
43901 CACCCCCACC CAGGCCACCT AAGGTCAATG TTGCTGTTAC AGTGAGCTTG
43951 TGGACCTGGA GATCCAGGTT GGGTTGAGCT GTGCCTGTGG CCCTCCTGCC
44001 TCCAGTCAGT GGGTGTTTGT TAGGTGCCTG CAGACCTCAG TACCGGGCAT
44051 GCTACAAGGA GCACACAGGG GAATGGCTCC TGCCTCCCTG GTGAACAGTC
44101 TCAGGGACTA ACCTCTCTCT TTCTCTCCTC CTCCTCCTCT TCTGCTGAGA
44151 ACTGGGAGGG GGGGTCAGGT AAGACGTGTG TCTCAGCTTG GGGGCAGCAG
44201 GGCTGGAGAG CTCACCCCCG ATCCACCCAG CTCCCTGGTG CATGTCTTTG
44251 GCACTGACCT TCCTGCCCCC AGACTTCTGT TCACTCAGGA GACTCACTTC
44301 TATGCCAAAT GACCAGAGCC CCTGCTTGGC TTGGCAGCAT CCCCTCCTGC
44351 CTTCTTCCCC ACTTCCCTTT TCTGGGTTCT TGCCTGTCCT CTGTGCATGC
44401 CCAGCTCTCC AGGAAAGAGG GTTTGCTTCC GTGTGAGTCC CATGTTGCTC
44451 CACGCTGCAT CTTCCACACA TGAACTCTGT CATTCTGACC CGGCTCAGTG
44501 TGCCCTCCAA GGGATGGGAT GGCCAGCTGC ATAGATTTTC TCAAACAGTT
44551 CTCCAGAACT TCCTCTGGTC TCAGCACCAT TAACAGTCAC CCTCCCTGTA
44601 GGTGACACAC AAAGCCACGG GCAAAGTGAT GGTCATGAAA GAGTTAATTC
44651 GATGTGATGA GGAGACCCAG AAAACTTTTC TGACTGAGGT AAGAAGATGG
44701 AGGGGGCCCG GGAGGTTGGT GTCACCATTG GAAGAGAGAA GACCTTACAA
44751 ATAATGGCTT CAAGAGAAAA TACAGTTTGG AATTACTGTC TTAAAGACTA
44801 AGCAGAAAAG AGCCCTAGAG GAATATCCCA CTCCCTCTAA ATTACAGCGT
44851 AATTATTTGT TCAATGAACA CTTACTAAAA GCAACACAAA CAGGGTACAA
44901 GGGATGCAGT AACAAAAGAT ACAGGGTTCA GAAGAGCTCT CAGGTTATGA
44951 GGATGATGGA CATGAAAACA CTCCAATTTA GTACAACTCA ATGTTATAAT
```

FIG.3-18

```
45001 CCTCACCTGA ACGCCCTGCT AAGGGAGCCT GGAGGGGAGC TCCCTGAGCA
45051 CTCACACTCC TTGGGCATTT ACAGTTTTCA CTACCCCTCC CAAGTTACTT
45101 CATGGAGTAA CTTAAGTTGG GGACACCTGT GGTCTGGGTA TTGCCCTCCA
45151 AGCCACTTGG CCACTCCCAC CCCAGTTCTC CCAATGCAGT TCCAAGGGTA
45201 AGGCCTATGA AGCCATCTCC ATCTATATGG TGGTGGTCTT CCCTCATCCT
45251 GATCTTAGTG CCCTGTCATA TCACAAGATA GGAGGTAGGA GATACAGGTG
45301 GTAACACTTG TCAAGCTGAT TCCTTGGAGG GAAGAGGTAA GGAAGACAGT
45351 GAGAAGTTAA CCACCAGCTT TCCTTGGCTT CCCCCACCCC CAGGTGAAAG
45401 TGATGCGCAG CCTGGACCAC CCCAATGTGC TCAAGTTCAT TGGTGTGCTG
45451 TACAAGGATA AGAAGCTGAA CCTGCTGACA GAGTACATTG AGGGGGGCAC
45501 ACTGAAGGAC TTTCTGCGCA GTATGGTGAG CACACCACCC CATAGTCTCC
45551 AGGAGCCTTG GTGGGTTGTC AGACACCTAT GCTATCACTA CCCTAGGAGC
45601 TTAAAGGGCA GAGGGGCCCT GCTTTGCCTC CAAAGGACCA TGCTGGGTGG
45651 GACTGAGCAT ACATAGGGAG GCTTCACTGG GAGACCACAT TGACCCATGG
45701 GGCCTGGACC ACGAGTGGGA CAGGGCTCAA CAGCCTCTGA AAATCATTCC
45751 CCATTCTGCA GGATCCGTTC CCCTGGCAGC AGAAGGTCAG GTTTGCCAAA
45801 GGAATCGCCT CCGGAATGGT GAGTCCCACC AACAAACCTG CCAGCAGGGC
45851 GAGAGTAGGG AGAGGTGTGA GAATTGTGGG CTTCACTGGA AGGTAGAGAC
45901 CCCTTCCTAT GCAACTTGTG TGGGCTGGGT CAGCAGCTAT TCATTGAGTT
45951 TGTCTGTGTC ACTGAAACTG ACCCCAGCCA ACTGTTCTCA GTTCACAGCC
46001 CTGTTTTCAA AGAATTACAC ATCTCTAAAG GCAAACAGGG CACGGACAAG
46051 GCAAACTGGA GAGGCAAACT GTAGCCTGAG ATGGCCTGGG CTTGCCATCA
46101 CAGGTATTCA GGTGCTGAGG GCCCTTAGAC CAACTAGAGC ACCTCACTGC
46151 CTAGGAAATC AATGAAGGGG AAATGAGTTC TAGCGGAGCC CTGAAGGATC
46201 AGAATTGGAT AAAGTTCTTA TTGGCAGAGA GGCACCAGGA TTGAAGTGAC
46251 AGGAGCAAAG ACCTGGGAGG AAAGAGGAGA AAATCATCTA TTTCACCTGG
46301 AAACAAATGA TTCCAAGCAT AGAAATAATA ACAGCTGACA AGTACTGAGT
46351 GCCCTCTATA TGCTAGGCAC TGGGCTGAGG GATTAACATG CATGTGCATG
46401 TTTATTCCTC ATGACAACCT TGGTTTCCAG ATAAGCTGGA CTGGAAAGGG
46451 ACAGAGCTGG GATCCTGGGC TAATCAGTCT GGTCGCCAAG CCTGAGACTT
46501 TAGCCACTGC CCTTCACATG GGGGTCCATG AAAATAGTAG TAGTCTGGAA
46551 CAGTTTGGGG GTACATCAAG GTCGCTGTGT TTTAAGCTAT GGAGTCTGGA
46601 CTATAGGAGA CAAATGTAAA AGAGTTTTTT GGTTGACTGG CTTTTTGGTT
46651 TTTTTGTTTG TTTGTTTGTT TGTTTGTTTG TTTGTTTGTT TTTTCCTGTT
46701 TCTGGGGCTT GAATCAGGAA GGAGGTTTTT TTGTTGTTGT TGTTTTGAGA
46751 AAGGATATTG CTCTGTTGCC CAGACTGGAG TGCAGTGGCA CGATCATGGC
46801 TCACTACAGC TTCGACCTCC TGGGCTCAAG CAATCCTCCT GCCTTAGCCT
46851 CCCAAGTAGC TGGACTACAG GTGTGTACCA CCACACCTAA TTTTTTGAAT
46901 TTTTTTTTCT TTTTTTTTTT TTTTTTTTTT GGTAGAGACA GGTTCTCACT
46951 TTGTTGCCCA GGCCTGAATC TCAAACTCCT GGGCTCAAGC ATTCCTCCTG
47001 CCTCGCCCTC CCAAAGTGTT GGGATTACAG TTGTGAGCCA CCATGCCCGG
47051 CAGGAAAAGA TTTTTAAGCA AGAAAGCTTA AGAGCTGTGG TTTTTCCAAA
47101 ATGAGTCTGG GCTGGCACAG TGGCTCATGC CTGTAATCCC AGCACTTTTT
47151 TGGGAGGCCG AGGTGAGTGG ATCACTTGAG GTCAGGAGTT TGAGACCAGC
47201 CTGGCCAACT GGTGAAACCC CTGTTTCTAC TAAAGAAAAA AATGCAAAAA
47251 TTAGCTGGGC GTGGTGGTGC ACGCCTGTAG TCCCAGCTAC TCAGGAGGCC
47301 GAGGCAGGAG AATAGCTTGA ACCTGGGAGG CAGAAGTTGC AGTGAGCCAA
47351 GATCACACCA CTGCATTCCA GCCTGGGTGA CAGAGTGAGA CTTCATCTCA
47401 AAAAAAAAAA AAAAGAGAGA CTGATATGGT TAGTACATTG GGGTGGAATG
47451 CGGAGGGTCC AGGGAATGGA GCCCTGCATA GGGGGCTAAT GAAACATTTC
```

FIG.3-19

```
47501 AGATTTCTGA ATTAAGGTAG TGGCTGTGGG GACAGGAGCC TGGGAGGCAG
47551 GGTGGAGTCA GAATGGAGAG ACTGGTTGGC AATGAGGGAA CAGGAGGAGG
47601 AGGAGGAGGA GTTACGAGTG GCTTGAGGTG TCACTTACCA GACATTTGGG
47651 GGATGGGGGA TAGCCGTGAT TGTTGAGCAA CTGGTTTGGG AAGAGCTAGC
47701 ATTGATCCCT GCTGTTCTGT GCTAGCAGAA CCTATCAGCA TCTTCTGGGC
47751 AGGAAACTGG CTCCATGAGA CTGGCTTAGG GAGAGGCTGC TAGTCACCTA
47801 ATCTGCAGAG AAGGGGCAGC TGGAGCTGTG GGACAGAAGA GGCATCCATG
47851 TAGCTGGTGG GGGTGTCTCA GCTTGTGAAG AGGAGATGGC TTTGAGCAGG
47901 GCTGACACTG AAAAGGCTGG AAGAAAAAAA CAGACACACA AGAGTCTCAG
47951 GATCAGGTAG CATAGGAAAG TTGTGGACAG TCTTTGAGGA GCACTCCCTC
48001 AGGCAGGCAG GCAGGCAGGT CATGAGCTAT AGCGATTCAG GAAGAGCTCC
48051 CTGGGTGTGT GAGCAGCTCC AGGAGCCTAA GGGATGAAAG TAGTATTGCA
48101 GGGGGCTGGA GAGCAAGGAG TGGCTCCTTC TACATTTGCA AGGGAAGGAG
48151 AAAGGAAGTT GCTCCTGAGA GTGGTAAGAG TCAGTGGTGG AGGCCTGGAG
48201 AGGAGACATA ACAAACAAAT TTGTTGACAA ACATTTTGGT AGGAAGGGGG
48251 AGAGCTTAAA GTTTAGACAG TGGGGAAGGT GGAGTCTTAG AGGAGGTGAA
48301 TGTCTGAAAG ACAGAGCTAG CTGGAGCAAG AAGTCACTTC TCTGTTGCAG
48351 GCAGGAAGGA TCCAAAGTGG CTCAAGCCAG AGATTGGGAG AGTGGGGAGG
48401 AGGGAGCAGC CTGGATCTAA GTAAAATGGG TAGAGGTGGA GGGGGTGCTG
48451 CAACGGCCAG GGTTTTCTGA AGTTGGGGAC ATTAGGAGAG AGCTGTGAGG
48501 GCTTTGGCCA GCCACTGTGC TAGTGATTGG TGAACCAAAG GATGGGCAGG
48551 AGATGGCAGC AGGGAAGCAG AGGAAGTCCA GGCTTCCTGT TGGTATTGGG
48601 ACAAGGGAGA GGCCATAGGA GGCCCTGGCC CTGTTGTCCA GGTTGGGTTC
48651 TGAAGCTGGG TGGGCATGGC CTGGTAGGAG AGCATCTATG GCGCCCAATT
48701 CCAGATTCAG GGTCTAGTTG ATTTGCTGGC CCTGTAGCCT CAGCTCATGC
48751 TTCTGTTCCA GGCCTATTTG CACTCTATGT GCATCATCCA CCGGGATCTG
48801 AACTCGCACA ACTGCCTCAT CAAGTTGGTA TGTCCCACTG CTCTGGGCCT
48851 GGCCTCCAGG GTCCTATCCT TCCTGGCTTC CTTGTCACAA AGGAGGCTGA
48901 CTTGTCCCCT CTGGCTAGAG GGCAGAGGTG TTGCCTAGGA GCTCCTATCT
48951 TTCCCTTCCT GCTTCTTCCA ATGCCCTTCT CTGTCCTCTG GGAGCTCCGA
49001 GACACACACA GACATAATTT CACCTTCTCT CATTAGCAAC CTTTGAAATA
49051 ATTTGATTAG AAGGGACTTC AGAAGTTTGT TGACTATATG TAGAAAACCC
49101 TGTCATTTTA CCTGCTTTTG CCCCATAGTA GTCTTGTAAA ACAGTTCATT
49151 GCTGACCCCA TTTTACAGTG GTGGCACCTG AAGCCTCAGC CTGAGGCCAC
49201 CGAGCTAGTA AATTTACAGG GACCAGTTTG AGACCAGCAT TCCTCCCACT
49251 GCCCCTCAGC TGTGGTGGTT ACAATGTTGT TTGTCTTACT GACTTGCTAT
49301 CTGGCTTCCT GGGTGTCTAC CGGCTGGCCC TGGCTCTGCC CTCTAGACCC
49351 ACACCACGCA ATCTTCATTC CTTTCCCACA TGACTGCCCT GTAGCTATTC
49401 AAAGAGCTTG TCTCCCCCAA GTCTCCCCAT CTACTGCCTC CACCTTGCCT
49451 TTTTCTGTCT TATCCTGGTT CTAGCCACTG CCTGAAATCA TTTTAGGAAT
49501 AAGACAGGAC AGGGAAAAAC AAAAGCAACC CCCTGTCCCA CCTCTGAGTT
49551 CCACTCTCCA AGTCCCTGAG CCTCACCTCC AGGGCTCCAG TGGCTCTGCC
49601 ATGAACCCAC TGTGGGCTGG GAGTCTGCTG TGCACAGATA CCAGACCCTC
49651 AGAAACACAA ATGCCAAGTG TGTCTGTTTT TTTGTTTTGT TTTGTTTTGT
49701 TTTTTAGATG GAGTCTCATT CTGTTTCCCA GGCTGGAGTG CAGTGGTGCA
49751 ATCTTGGCTT ACTGCAGCCT CTACCTCCCG GGTTCTAGTG ATTGTTCTGC
49801 TTCAGCCTCC CAGTAGCTAG GACTACAGGC GTGTGCCACC ACGCCCAGCT
49851 AATTTTTTTT TTTTTTTTTT TGTATTTTTA GTAGAGACAG GGTTTTGCCA
49901 TGTTGGCCAG GCTGGTCTTG AACTCCTGAC CTCAGGTGAT TCACCCGCCT
49951 TGGCCTCCCA AAGTTCTGGG ATTACAGGTG GAAGCCACCG TGCCTGGCCT
```

FIG.3-20

```
50001 GAGTGTGTCT ATTTGATAGA GCTTTCTGCT CTGATTCTCC CTTGCTATAC
50051 ACCTTTTCTC CCCTTCTCAG TGGCTTCTCT TGCCTATGCT TCCTCCCCAG
50101 GGCCAGGTTT GAGAACATCC CCATGAAGTC CTGACCTGTC TTTTATCCTA
50151 CCAGGACAAG ACTGTGGTGG TGGCAGACTT TGGGCTGTCA CGGCTCATAG
50201 TGGAAGAGAG GAAAAGGGCC CCCATGGAGA AGGCCACCAC CAAGAAACGC
50251 ACCTTGCGCA AGAACGACCG CAAGAAGCGC TACACGGTGG TGGGAAACCC
50301 CTACTGGATG GCCCCTGAGA TGCTGAACGG TGAGTCCTGA AGCCCTGGAG
50351 GGGACACCCG CAGAGGGAGG ACAGATGCTG CCCTTGCATC AGAGCCCTGG
50401 GAATTCCAGG GGAGGCCTGT GAAGCGTAGG ACCGGATACC CAGAGCTGAG
50451 GATATTTTTC CCTTGCCAGG TGGGGCCTCA CGATTTAGCT CCTGAGCTCA
50501 GGGGGCTGGG AACTGATCAG TGTCCCATCA TGGGGGATAA GGTGAGTTCT
50551 GACTGTGGCA TTTGTGCCTC AGGGATCGCT AAGAGCTCAG GCTATTGTCC
50601 CAGCTTTAGC CTTCTCTCTC CATGGTGAGA ACTGAAGTGT GGTGCCCTCT
50651 GGTGGATAAT GCTCAAACCA ACCAGAGATG CTGGTTGGGA TTCTTGAAAT
50701 CAGGGTTGTG AGGCCTCAGA AATGGTCTGA ATACAATCCA TTTTGGAGTC
50751 TGAGGCCCAG AGAAGTTCAG TGAATTGCCT AGGAGCATAC AGCTGCCTAA
50801 TGGCAGAGGC TAGATGAACC CTAGTCTGGT TCTTTTCCAC TTTAACGTGC
50851 AGTTTCATCC TAGGCAGTGT TATGTTATAA GGGCTCTCCA AGGCAGTTCA
50901 CCTACGGCTG AGGAAGGACT ATTTTCAGGT GGTGTCTGCG CAGGACAGCC
50951 TGTGGGGTGT CCCTACAGAA CCTGTTCTAG CCCTAGTTCT TAGCTGTGGC
51001 TTAGATTGAC CCTAGACCCA GTGCAGAGCA GGTAAGGGAT GTAAACTTAA
51051 CAGTGTGCTC TCCTGTGTTC CCCAAGGAAA GAGCTATGAT GAGACGGTGG
51101 ATATCTTCTC CTTTGGGATC GTTCTCTGTG AGGTGAGCTC TGGCACCAAG
51151 GCCATGCCCG AGGCAGCAGG CCTAGCAGCT CTGCCTTCCC TCGGAACTGG
51201 GGCATCTCCT CCTAGGGATG ACTAGCTTGA CTAAAATCAA CATGGGTGTA
51251 GGGTTTTATG GTTTATAACG CATCTGCACA TCTTTGCCAC GTTCGTGTTT
51301 CATTGGTCTT AAGAGAAGGA CTGGCAGGGT TTTTTTGTTT TAGATGGAGC
51351 CTCACTTCGT TGCCCAGGCT GGAGTGCAGT GGCACAATCT GGGCTCACTG
51401 CAACCTCTGC CTTCTGGGTT CAAGTGATTC TCCTGCCTCA GCCTCCCAAG
51451 TAGCTGGGAC TACCGGCACA CACCACCATG CCCGGCTAAT TTTTGTATTT
51501 TTAGTAGAGA CAGGGTTTCA CCATGTTGGC CAGGCTGGTC TTGAACTCCG
51551 GACCTCAGGT GATCCGCCTG CCTCAGCCTC TAAAAGTGCT GGAATTAATA
51601 GGCGTGAGCT ACCTCGCCCG GCCAGGTTTT TTTTTTTTTT TTTTTAGTTG
51651 AGGAAACTGA GGCTTGGAAG AGGGCAGTGG CTTGCACATG GTCGATAAGG
51701 GGCAGATGAG ACTCAGAATT CCAGAAGGAA GGGCAAGAGA CTGTTCATGT
51751 GGCTGTCTAG CTAGCTCTTG GGCCAAATGT AGCCCTTCTC AGTTCCCTTC
51801 AAGTAGAAGT AGCCACTCTA GGAAGTGTCA GCCCTGTGCC AGGTACCACG
51851 TGGACAGAGT GAGGAATCTT GGAAAGATTC CTACCTTTAG GAGTTTAGTC
51901 AGGTGACAGC ATATCTCAGC GACTCAAACA CACACACATT CAAAGCCTTC
51951 TGTAATTCCT ACAAAGTTGT GAGGGGTAGA GGAGAGGAGA GACAAGGGAT
52001 GGTTAGGATA ATGAAGGAAT GTTTTGTTTT TGTTTTTGTT TTTGAGATGG
52051 AGTTTCACTC TGTCACCCAG GCTGGAGTGC AGAGGTGCAA TCTTGGCTCA
52101 CTGCAGCCTC CGCCTCCCAG GTTCAAGCAA TCCTCCTGCC TCAGCCTCCC
52151 AAGTAGCTGG GACTACAGGT GTGCGCCACC ACGCCTGGCT AATTTTTGTA
52201 TTTTCAGTAG AGACAGGGTT TCGCCATATT GGCCAGGCTG GTCTCAAATG
52251 CCTGACCTCA GGTGATACAC CCGCTTCAGC CTCCCAAAGT GCTGAGATTA
52301 CAGGCATGAG CTACCGTGCC TGGCCATGAA GGAAGATTTG TTTTAAAAAA
52351 TTGTTTTCTT TAATATTAAT TGAACACCTC TGTTCAGAGC ACTGGGCTGG
52401 TGCCAGAGGG TTTCAGACAT GAATCAGATC CAGCACCTCA TAGAGCCTTA
52451 ATCTGGCACA CACACACAGC CACAAGGAGA CACAGACAAG GCAGGGTAGG
```

FIG.3-21

```
52501 ATGAGTGGAA GCTAGGAGCA GATGCTGATT TGGAACACTT GGCTTCTGCA
52551 GTGAAGCCCC TTCTTAGTCC TCTTCAGTAA CCCAGCTCTC AGTGGATACA
52601 GGTCTGGATT AGTAAGATTT GGAGAGATGA TTGGGGATTG GGGAGAGCTC
52651 TCTAACCTAT TTTACCACCT CCTCTTCTGC CATTCTTCCT GTCCACATCC
52701 CCAGCATCCC TTTCCCTTGC CAAGTATCTG TGGCCTCTGT AGTCCTTTGT
52751 AAACAGCTGT CTTCTTACCC TACAGATCAT TGGGCAGGTG TATGCAGATC
52801 CTGACTGCCT TCCCCGAACA CTGGACTTTG GCCTCAACGT GAAGCTTTTC
52851 TGGGAGAAGT TTGTTCCCAC AGATTGTCCC CCGGCCTTCT TCCCGCTGGC
52901 CGCCATCTGC TGCAGACTGG AGCCTGAGAG CAGGTTGGTA TCCTGCCTTT
52951 TTCTCCCAGC TCACAGGGTC CTGGGACGTT TGCCTCTGTC TAAGGCCACC
53001 CCTGAGCCCT CTGCAAGCAC AGGGGTGAGA GAAGCCTTGA GGTCAAGAAT
53051 GTGGCTGTCA ACCCCTGAGC CATCTGACAA CACATATGTA CAGGTTGGAG
53101 AAGAGAGAGG TAAAGACATA GCAGCAAGTA ATCTGGATAG GACACAGAAA
53151 CACAGCCATT AAAAGAAAGT TTAAAAGAAG GAAATTCACC CAAACCATTT
53201 GAATACAGTA AGTGTATTCA TCTTTCGATA TTCCCCTGTC CATATCTACA
53251 CATATACTTT TTTTTATAGT AAATAGTTCT GTATTTTGCC CTGCATTTCC
53301 CTTGTGTTTA CTATCCAGTC TTCCTGTTTA TCATTTTTGT CGACAACATG
53351 AAATTCTATT GAGAGACTGT CTGAACATAT TGTAATGTAG ATGTTCAGGT
53401 TTTTCCAGTT TCTCTTTACA ATAGGTATTT AACTACAGTG AGCAGTTTTA
53451 TGCATTTAGC TAATTTCTCC TTTGAGGAAG TATTTTCAAA ATTACCTTTA
53501 TTCTTCTCAG GTAATAATTT CATTATTACC AAAGTTACCC TAGGTCTTTT
53551 CAAGTGTGTG GTTAAAAAAC GAGAATCTGG CTGGGCGCGA TGGCTCACAC
53601 CTGTAATCCC AGCACTTTGG GAGGCTGAGG CTGGTGGATC ACCTGAGGTC
53651 TGGAGTTCGA GACCAGCCTG GCCAACATGG TGAAACCCCA TCTCTACTAA
53701 AAATACAAAA CTTAGCCAGG CATGGTGGCA GGTGCCTGTA ACCCCAGCTA
53751 CTTGGGAGGC TGAGGCAGGA GAATTGCTTG AACCCAGGGG CGGAGGTTGC
53801 AGTGAGCCGA TATCACGCCA TTGCACTCCA GCCTCGGCAA CAAGAGTGAA
53851 ACTCTGTCTC AAAAATGGGG TTCTTTTCCT GCCATCAAAA ATCATGTTTC
53901 TTTTAAAAAC AAGTTCAAAC ATTACCAAAG TTTATAGCAC AGGAAATACG
53951 TCTTCTGTAA TCTCCCTTAA CCAATATATC CCTCAACATT CTCCTCACCC
54001 CCAACTCCAC CCTCCCAGGA TAACCAGTTG GGACATAATC TTTATTTAAA
54051 AATGGTTTCC GGATAGAGAA AGCGCTTCGG CGGCGGCAGC CCCGGCGGCG
54101 GCCGCAGGGG ACAAAGGGCG GGCGGATCGG CGGGGAGGGG GCGGGGCGCG
54151 ACCAGGCCAG GCCCGGGGGC TCCGCATGCT GCAGCTGCCT CTCGGGCGCC
54201 CCCGCCGCCG CCCTCGCCGC GGAGCCGGCG AGCTAACCTG AGCCAGCCGG
54251 CGGGCGTCAC GGAGGCGGCG GCACAAGGAG GGGCCCCACG CGCGCACGTG
54301 GCCCCGGAGG CCGCCGTGGC GGACAGCGGC ACCGCGGGGG GCGCGGCGTT
54351 GGCGGCCCCG GCCCCGGCCC CCAGGCCAGG CAGTGGCGGC CAAGGACCAC
54401 GCATCTACTT TCAGAGCCCC CCCCGGGGCC GCAGGAGAGG GCCCGGGCTG
54451 GGCGGATGAT GAGGGCCCAG TGAGGCGCCA AGGGAAGGTC ACCATCAAGT
54501 ATGACCCCAA GGAGCTACGG AAGCACCTCA ACCTAGAGGA GTGGATCCTG
54551 GAGCAGCTCA CGCGCCTCTA CGACTGCCAG GAAGAGGAGA TCTCAGAACT
54601 AGAGATTGAC GTGGATGAGC TCCTGGACAT GGAGAGTGAC GATGCCTGGG
54651 CTTCCAGGGT CAAGGAGCTG CTGGTTGACT GTTACAAACC CACAGAGGCC
54701 TTCATCTCTG GCCTGCTGGA CAAGATCCGG GCCATGCAGA AGCTGAGCAC
54751 ACCCCAGAAG AAGTGAGGGT CCCCGACCCA GGCGAACGGT GGCTCCCATA
54801 GGACAATCGC TACCCCCCGA CCTCGTAGCA ACAGCAATAC CGGGGGACCC
54851 TGCGGCCAGG CCTGGTTCCA TGAGCAGGGC TCCTCGTGCC CCTGGCCCAG
54901 GGGTCTCTTC CCCTGCCCCC TCAGTTTTCC ACTTTTGGAT TTTTTTATTG
54951 TTATTAAACT GATGGGACTT TGTGTTTTTA TATTGACTCT GCGGCACGGG
```

FIG.3-22

```
55001 CCCTTTAATA AAGCGAGGTA GGGTACGCCT TTGGTGCAGC TCAAAAAAAA
55051 AAAAAAAAAT GATTTCCAGC GGTCCACATT AGAGTTGAAA TTTTCTGGTG
55101 GGAGAATCTA TACCTTGTTC CTTTATAGGC CAAGGACCGC AGTCCTTCAG
55151 TAACACCAGT GTAAAAGCTT GAGGAGAAAT TGTGAAGCTA CACAGTATTT
55201 GTTTTCTAAT ACCTCTTGTC ATTCTAAATA TCTTTAATTT ATTAAAAAAT
55251 ATATATATAC AGTATTGAAT GCCTACTGTG TGCTAGGTAC AGTTCTAAAC
55301 ACTTGGGTTA CAGCAGCGAA CAAAATAAAG GTGCTTACCC TCATAGAACA
55351 TAGATTCTAG CATGGTATCT ACTGTATCAT ACAGTAGATA CAATAAGTAA
55401 ACTATATTGA ATATTAGAAT GTGGCAGATG CTATGGAAAA AGAGTCAAGA
55451 CAAGTAAAGA CGATTGTTCA GGGTACCAGT TGCAATTTTA AATATGGTCG
55501 TCAGAGCAGG CCTCACTGAG GTGACATGAC ATTTAAGCAT AAACATGGAG
55551 GAGGAGGAGT AAGCCTGAGC TGTCTTAGGC TTCCGGGGCA GCCAAGCCAT
55601 TTCCGTGGCA CTAGGAGCCT GGTGTTTCCG ATTCCACCTT TGATAACTGC
55651 ATTTTCTCTA AGATATGGGA GGGAAGTTTT TCTCCTATTG TTTTTAAGTA
55701 TTAACTCCAG CTAGTCCAGC CTTGTTATAG TGTTACCTAA TCTTTATAGC
55751 AAATATATGA GGTACCGGTA ACATTATGCC CATTTCTCAC AGAGGCACTA
55801 CTAGGTGAAG GAGTTTGCCT GACGTTATAC AACCAGGAAG TAGCTGAGCC
55851 TAGATCCCTT CCACCCACCC CATGGCCCTG CTCATGTTCC ACCTGCCTCT
55901 AATTTACCTC TTTTCCTTCT AGACCAGCAT TCTCGAAATT GGAGGACTCC
55951 TTTGAGGCCC TCTCCCTGTA CCTGGGGGAG CTGGGCATCC CGCTGCCTGC
56001 AGAGCTGGAG GAGTTGGACC ACACTGTGAG CATGCAGTAC GGCCTGACCC
56051 GGGACTCACC TCCCTAGCCC TGGCCCAGCC CCCTGCAGGG GGGTGTTCTA
56101 CAGCCAGCAT TGCCCCTCTG TGCCCCATTC CTGCTGTGAG CAGGGCCGTC
56151 CGGGCTTCCT GTGGATTGGC GGAATGTTTA GAAGCAGAAC AAGCCATTCC
56201 TATTACCTCC CCAGGAGGCA AGTGGGCGCA GCACCAGGGA AATGTATCTC
56251 CACAGGTTCT GGGGCCTAGT TACTGTCTGT AAATCCAATA CTTGCCTGAA
56301 AGCTGTGAAG AAGAAAAAAA CCCCTGGCCT TTGGGCCAGG AGGAATCTGT
56351 TACTCGAATC CACCCAGGAA CTCCCTGGCA GTGGATTGTG GGAGGCTCTT
56401 GCTTACACTA ATCAGCGTGA CCTGGACCTG CTGGGCAGGA TCCCAGGGTG
56451 AACCTGCCTG TGAACTCTGA AGTCACTAGT CCAGCTGGGT GCAGGAGGAC
56501 TTCAAGTGTG TGGACGAAAG AAAGACTGAT GGCTCAAAGG GTGTGAAAAA
56551 GTCAGTGATG CTCCCCCTTT CTACTCCAGA TCCTGTCCTT CCTGGAGCAA
56601 GGTTGAGGGA GTAGGTTTTG AAGAGTCCCT TAATATGTGG TGGAACAGGC
56651 CAGGAGTTAG AGAAAGGGCT GGCTTCTGTT TACCTGCTCA CTGGCTCTAG
56701 CCAGCCCAGG GACCACATCA ATGTGAGAGG AAGCCTCCAC CTCATGTTTT
56751 CAAACTTAAT ACTGGAGACT GGCTGAGAAC TTACGGACAA CATCCTTTCT
56801 GTCTGAAACA AACAGTCACA AGCACAGGAA GAGGCTGGGG GACTAGAAAG
56851 AGGCCCTGCC CTCTAGAAAG CTCAGATCTT GGCTTCTGTT ACTCATACTC
56901 GGGTGGGCTC CTTAGTCAGA TGCCTAAAAC ATTTTGCCTA AAGCTCGATG
56951 GGTTCTGGAG GACAGTGTGG CTTGTCACAG GCCTAGAGTC TGAGGGAGGG
57001 GAGTGGGAGT CTCAGCAATC TCTTGGTCTT GGCTTCATGG CAACCACTGC
57051 TCACCCTTCA ACATGCCTGG TTTAGGCAGC AGCTTGGGCT GGGAAGAGGT
57101 GGTGGCAGAG TCTCAAAGCT GAGATGCTGA GAGAGATAGC TCCCTGAGCT
57151 GGGCCATCTG ACTTCTACCT CCCATGTTTG CTCTCCCAAC TCATTAGCTC
57201 CTGGGCAGCA TCCTCCTGAG CCACATGTGC AGGTACTGGA AAACCTCCAT
57251 CTTGGCTCCC AGAGCTCTAG GAACTCTTCA TCACAACTAG ATTTGCCTCT
57301 TCTAAGTGTC TATGAGCTTG CACCATATTT AATAAATTGG GAATGGGTTT
57351 GGGGTATTAA TGCAATGTGT GGTGGTTGTA TTGGAGCAGG GGGAATTGAT
57401 AAAGGAGAGT GGTTGCTGTT AATATTATCT TATCTATTGG GTGGTATGTG
57451 AAATATTGTA CATAGACCTG ATGAGTTGTG GGACCAGATG TCATCTCTGG
```

FIG.3-23

```
57501 TCAGAGTTTA CTTGCTATAT AGACTGTACT TATGTGTGAA GTTTGCAAGC
57551 TTGCTTTAGG GCTGAGCCCT GGACTCCCAG CAGCAGCACA GTTCAGCATT
57601 GTGTGGCTGG TTGTTTCCTG GCTGTCCCCA GCAAGTGTAG GAGTGGTGGG
57651 CCTGAACTGG GCCATTGATC AGACTAAATA AATTAAGCAG TTAACATAAC
57701 TGGCAATATG GAGAGTGAAA ACATGATTGG CTCAGGGACA TAAATGTAGA
57751 GGGTCTGCTA GCCACCTTCT GGCCTAGCCC ACACAAACTC CCCATAGCAG
57801 AGAGTTTTCA TGCACCCAAG TCTAAAACCC TCAAGCAGAC ACCCATCTGC
57851 TCTAGAGAAT ATGTACATCC CACCTGAGGC AGCCCCTTCC TTGCAGCAGG
57901 TGTGACTGAC TATGACCTTT TCCTGGCCTG GCTCTCACAT GCCAGCTGAG
57951 TCATTCCTTA GGAGCCCTAC CCTTTCATCC TCTCTATATG AATACTTCCA
58001 TAGCCTGGGT ATCCTGGCTT GCTTTCCTCA GTGCTGGGTG CCACCTTTGC
58051 AATGGGAAGA AATGAATGCA AGTCACCCCA CCCCTTGTGT TTCCTTACAA
58101 GTGCTTGAGA GGAGAAGACC AGTTTCTTCT TGCTTCTGCA TGTGGGGGAT
58151 GTCGTAGAAG AGTGACCATT GGGAAGGACA ATGCTATCTG GTTAGTGGGG
58201 CCTTGGGCAC AATATAAATC TGTAAACCCA AAGGTGTTTT CTCCCAGGCA
58251 CTCTCAAAGC TTGAAGAATC CAACTTAAGG ACAGAATATG GTTCCCGAAA
58301 AAAACTGATG ATCTGGAGTA CGCATTGCTG GCAGAACCAC AGAGCAATGG
58351 CTGGGCATGG GCAGAGGTCA TCTGGGTGTT CCTGAGGCTG ATAACCTGTG
58401 GCTGAAATCC CTTGCTAAAA GTCCAGGAGA CACTCCTGTT GGTATCTTTT
58451 CTTCTGGAGT CATAGTAGTC ACCTTGCAGG GAACTTCCTC AGCCCAGGGC
58501 TGCTGCAGGC AGCCCAGTGA CCCTTCCTCC TCTGCAGTTA TTCCCCCTTT
58551 GGCTGCTGCA GCACCACCCC CGTCACCCAC CACCCAACCC CTGCCGCACT
58601 CCAGCCTTTA ACAAGGGCTG TCTAGATATT CATTTTAACT ACCTCCACCT
58651 TGGAAACAAT TGCTGAAGGG GAGAGGATTT GCAATGACCA ACCACCTTGT
58701 TGGGACGCCT GCACACCTGT CTTTCCTGCT TCAACCTGAA AGATTCCTGA
58751 TGATGATAAT CTGGACACAG AAGCCGGGCA CGGTGGCTCT AGCCTGTAAT
58801 CTCAGCACTT TGGGAGGCCT CAGCAGGTGG ATCACCTGAG ATCAAGAGTT
58851 TGAGAACAGC CTGACCAACA TGGTGAAACC CCGTCTCTAC TAAAAATACA
58901 AAAATTAGCC AGGTGTGGTG GCACATACCT GTAATCCCAG CTACTCTGGA
58951 GGCTGAGGCA GGAGAATCGC TTGAACCCAC AAGGCAGAGG TTGCAGTGAG
59001 GCGAGATCAT GCCATTGCAC TCCAGCCTGT GCAACAAGAG CCAAACTCCA
59051 TCTCAAAAAA AAAAA  (SEQ ID NO:3)
```

FEATURES:

Start:	3000
Exon:	3000-3044
Intron:	3045-45393
Exon:	45394-45525
Intron:	45526-45761
Exon:	45762-45818
Intron:	45819-50154
Exon:	50155-50329
Intron:	50330-51076
Exon:	51077-51132
Intron:	51133-52775
Exon:	52776-52933
Intron:	52934-55922
Exon:	55923-56064
Stop:	56065

FIG.3-24

CHROMOSOME MAP POSITION:
Chromosome 22

ALLELIC VARIANTS (SNPs):

DNA Position	Major	Minor	Domain
941	A	T	Beyond ORF(5')
2612	G	A	Beyond ORF(5')
5080	G	A	Intron
6599	-	A C	Intron
6983	C	G	Intron
9885	A	-	Intron
12538	G	T	Intron
17707	T	C	Intron
18219	-	A	Intron
19670	C	T	Intron
21153	G	T	Intron
24566	C	-	Intron
26604	G	A	Intron
27255	C	G	Intron
27399	T	C	Intron
28088	G	A	Intron
28734	G	A	Intron
29246	-	T	Intron
29490	G	A	Intron
29934	T	C	Intron
34480	A	G	Intron
38812	T	C	Intron
40731	C	G	Intron
41303	T	A	Intron
41305	-	A	Intron
41457	G	C	Intron
43168	A	- T	Intron
43357	T	G	Intron
45664	T	C	Intron
47549	A	C	Intron
47908	C	A	Intron
52267	C	A	Intron
54654	T	C	Intron
54679	C	G	Intron
54693	A	C	Intron
54706	T	C	Intron
54712	T	C	Intron
54799	T	C	Intron
54819	G	A	Intron
55499	C	T	Intron
56825	C	A	Beyond ORF(3')
58871	T	A	Beyond ORF(3')

Context:

FIG.3-25

```
DNA
Position
941       GAGTAAGTGGGTGGTCAGGTTACAGACTTAATTTTGGGTTAAAAAGTAAAAACAAGAAAC
          AAGGTGTGGCTCTAAAATAATGAGATGTGCTGGGGGTGGGGCATGGCAGCTCATAAACTG
          ACCCTGAAAGCTCTTACATGTAAGAGTTCCAAAAATATTTCCAAAACTTGGAAGATTCAT
          TTGGATGTTTGTGTTCATTAAAATCTCTCACTAATTCATTGTCTTGTCCACTGTCCGTAA
          CCCAACCTGGGATTGGTTTGAGTGAGTCTCTCAGACTTTCTGCCTTGGAGTTTGTGAGAG
          [A,T]
          GATGGCATACTCTGTGACCACTGTCACCCTAAAACCAAAAAGGCCCCTCTTGACAAGGAG
          TCTGAGGATTTTAGACCCAGGAAGAATGAGTGATGGGCATATATATATCCTATTACTGAG
          GCATGAGAAGAGTGGAATGGGTGGGTTGAGGTGGTGTTTTAAGGCCTCTTGCCAGCTTGT
          TTAACTCTTCTCTGGGGAACGAGGGGGACAACTGTGTACATTGGCTGCTCCAGAATGATG
          TTGAGCAATCTTGAAGTGCCAGGAGCTGTGCTTTGTCTATTCATGGCCCCTGTGCCTGTG

2612      TGAGTTGGAACAGTTTGATACCAAAACCATCCCCCCGCCCCCCAACCCCCAGCCTAGGGT
          CCGTGGAAAAATTGGCCCCTGGTGCCAAAAAGGTTGAGGACTGCTGATCTAGAGGACCAA
          TTTATTCAATGTTGGTTGAGTAAATGAGCTCTTGGATTAGGTGATGGAAAAATCTGAAAA
          AACAGGGCTTTTGAGGAATAGGAAAAGGCAGTAACATGTTTAACCCAGAGAGAAGTTTCT
          GGCTGTTGGCTGGGAATAGTCATAGGAAGGGCTGACACTGAAAAGAAGGAGATTGTGTTC
          [G,A]
          TTTCTTCTTCTCAGAGCTATAAGCAAAGGCTGAAAGTTCTAGAAAAAGGCAAGTTTTGTT
          TCAGTAGAAAAAAGGATAATCAGAACCATTTTTAGAAAATGGAATGAGACTACTTTTGAG
          GCCATGAGTTCCTTGTCCCTGGAGAGATGAGCAGAGGTTGGACAAGTGCTTACCAGAGAT
          CTTGTGGAGGCAGAAACTGTGCATCTAGCAGAGCATTGGCCTAACCCTTTCAAATGAGAT
          GCTGTTAACTCAGTCTTATTCTACATGGTAGGAATCCTGTCCCTTTGCCTCCTGCTACTT

5080      ACAACGTAAAATAGTTGAAATTTGTTGGTGGAAAGAAGAGCAGTCCACTCCAGAGGCTGG
          ATGGGCATGCCTGGCCCCCAAGGTCTGAAGTGGTAGGGCTGTGCCTATATCCTGAGAATG
          AGATAGACTAGGCAGGCACCTTGTGCTGTAGATTCCAGCTCCTGCACATAGCTCTTGTTG
          TAAAACATCCCTGTGCTTATACCAAGTAATTGAGTTGACCTTTAAACACTTGCCTCTTCC
          CTGGGAACCATATAGGGGATTGGCCTGGAGACGTCTGGCCTCTGGAAGAGTTGGAAAGCA
          [G,A]
          CCATCATTATTATCCTTTCCTTTCAGCTATAACTCAGAGCTCTCAAGTCTTTTCTGTGGA
          TCTTATTGCCTTGGTTCTTGCCCCTTTTACTCCCAGGGAAGTTGATTCTGTCTTTTCTGT
          TCCATTTAGTATGACAGGAGCAGAGAATGTCAGAGCTGTAAGGGACCTTATAGTTAAAGC
          CTTTGGCTGGTCCTTTCATTTTATAGCTGGGACTAATAAGTAACGTCAAAACCCAATGAG
          TTCACAGATTGGGTCTCGCCTTGGCATGTAACCCATATGTTCATATTCTTGCTGTTTTCC

6599      CTGTAATCCTAGCACTCTGGGAGGCCGAGGCAGAAGGATCGCTTGAGCCCATGAGCCCAG
          GAGTTTGAGACCAGCCTGGCCAACATGGCAAAACTCCACCTCTACAAAAAATACAAAAAT
          ATTAGCCAGGCGTGATGGCACACACCTGTAGTCCCAGCTACTTGGGAAGCTGAGGAGCGA
          TGATTACCTGAGCCCAGGGATATCAAGGCTGTAGTGAGCTGTGATCATGCCACTGTACTC
          CATCCAGCTGGGGGACAGAGTGAAACCCCTGTCTCAAAACAAAACAAATGAAAAAAAAAA
          [-,A,C]
          CCTTAATAATCAGTAACTGTCACTTTATATTATGTTGTGAGTGTGTGTCTATATACACCT
          ATATGTATACATTTCTCTTATTACACATTCATTGGTGATCTGATGTGGAGCCCCAGGGAT
          TAAGGGCAACTTTGAACTACCCTGACACAATCAAGCCAAATATCATTCCCGTGGAGGAAG
          TAGAGTATCTAGGTTCTGTCTCCTAGTTGCAGCTTTACCTTGAGGACAGAGACTCTAATC
          CAGCTGTGCTGAAGGAGCACATCTCCTGACTTCTGAGCTTTCCCCTGGTAAATTCAAACT
```

FIG.3-26

```
6983    CACATTCATTGGTGATCTGATGTGGAGCCCCAGGGATTAAGGGCAACTTTGAACTACCCT
        GACACAATCAAGCCAAATATCATTCCCGTGGAGGAAGTAGAGTATCTAGGTTCTGTCTCC
        TAGTTGCAGCTTTACCTTGAGGACAGAGACTCTAATCCAGCTGTGCTGAAGGAGCACATC
        TCCTGACTTCTGAGCTTTCCCCTGGTAAATTCAAACTGGATGTCACGGCGCCCTCAGATA
        GAGCCTGGTAATTTGCCCTGGGGAGAGTGACTGTCTTTTGGATCTAATTTGACTTTTGCC
        [C,G]
        CAGTTGGAGGAAAATCTTCAGGGCTAGGAAGGATTGTATTTGTCTGACCCCAGAGATAAC
        CTGGGTTTTGAGGAACATGGGGCATCAACCTGAATGGTCTTGTAAGATCTCTCCCACGCC
        AGCTTGCCAGTGTTTCTCTGATGAATTTAGAGTACCTGAGTAGTGCAGGCCTGCTGGGAG
        GAGGACTCTCCCTCTGTGCTACTCAGAGAAATTCATTCTTCAAGGCCCCCTTCCAGCCTT
        GCTCTTACCCAGCTGGGCTACAGTTACAATAAAGGAAATGACTTTTCTTCTCCCCTTCCC

9885    GGCGTGCCACCACACCTTGCCATTTTTTTTTATTTTAAGTAGAAACAAGGTCTTATTAAT
        ACTATGTTGCCCAGGCTGGTCTTGAACTCCAGCGATCCTCCTGCCCCAGCCTCCCAAAGT
        GCTTGGGATTACGGAAGTAAGCCACTGTGCCTGGCCAGTGCAACCCCCATTTTATACTAA
        AACAGGAAGGCCCAGAAAGGTTTGGAGTAACTTGTCCAGGGTCACACAGATGATATTTGA
        ACTCAGGTCTCCCTGGCTCCCAAGAGAGTCTGCTTTCCACTAGGACTCCCAGGAGAAAAA
        [A,-]
        AAAAAAAAAAACAGTAGACTTGGAGACAGAAAATCTGATTTGAGTCTTAGTTGAGCTAGG
        CTAACTGTGTAACTGTGGGCAAGTTCCTTAGCCCCTGTGAGCCTCAGTTTCTTATCTGTA
        AAATGTCATAAAAGAAATCCATCTCATGGAGTAGTTGTGATGATCAAGGACTCTGAAAAC
        ATTAGAATGGTTTAATGTGAAGGATTAGCAGCAGCACATGGCAACATTGTGCATCTTATA
        TTAACTATCCAAATATATCAAGCGTCATTTGCTATATATAAAAGTCATCAAATTAGGCAC

12538   ACTTGGGAGGCTGAGGCAGGAGAATCACTTGAACCTGGGAGGCAGAGGTTGCAGTGAGCC
        CAGATCACGCCACTGCACTCCAGCCTGGTGACAGAGTAAGACTCCATCTCAAAAAAAAAA
        AAAAAAAAAAAAATTCCTTAATTTGGCCTACAGTAGAGCCCTCCGTAATGTGGCCTCTCT
        CCACATCTCCACAACCTCCTGCTCCCTGCACTTCAGCCTCACCTCTCTTCTGGACAGGCC
        CTCCTTCTGACAAGGGCTTTGTTCATTCTGCTCCCTCTGCCTAGAATGCCCCCTTACTCT
        [G,T]
        TTCACTTAACTCCTGCTTATCGTTTAGATCTTTACCTGGATGGCTCAGAGAAATATAGAA
        GTAATTCCTCACCCTGAAAAATAGGTTAGGTCCCTGTTTTATGTTTTCATAGACCTTTCC
        TTTGAGGCTTTTTTTAAAAAAGTAGTTTTAATCTCACATTTATTCATGTGATCATCTCCT
        TAATGATATCTTAAGACCTCTAATAGAACAATTTGGTCATGGACTGTGGGGTTTTTGCCC
        CTCATTGTGTCAGCACTGAGCATATTGTTGGCATAGGAGGGATATTTGTTGAATGAATTG

17707   GTAGTGGGTGCTCAGAGTGTTTGCTGGGTGAATGATGTATTTGTTGAACGACTCTTTGGA
        CACTTGAATAAAGTCCATCCAGTATGCACCATTACCATCTCTTCGCTCTACAATATTCTT
        TTAGGCAAGAGCTTATCTTTTGAGGTGATAAGATAAGCTCAAACTTATGTAGACTAAGAC
        CTCAGTCTGTAAATGTCATCCCTAAGTCTTAAACCATCAAAACCAGGGCCTCAAGGAATG
        GCATGCCTTCTGCAACTGTAGCAACCTGCTGTGCTTATTTTGCCGTGTTTTTCATTTTTC
        [T,C]
        CCCAAAAGCTAGAGTCCCTTCTCCCATGGGCAGTGCTGGAAGTGTGCTAACAAATTCTTT
        CTCCATACTGCTTACGATTACAAAAAAAACCCTCAGCATCTCATGCCAGACTTGAGTTAA
        GGTTGTTTTCTTTTGTGTGTCAGCTGTATTCTGGTCATGACTTCCTGATGATGCCCTATA
        GAGATTTTGCTGAGATCAGAGGGTGCTCCACTGCCATCAGTAGCACTGACTCTTGCAGAA
        GCACCGTTTCTGAAGTTGGCTAATGTCATCCCTCACGTTTGTTTGTTTGAAATTTGTTTT
```

FIG.3-27

```
18219    TGCCATCAGTAGCACTGACTCTTGCAGAAGCACCGTTTCTGAAGTTGGCTAATGTCATCC
         CTCACGTTTGTTTGTTTGAAATTTGTTTTAGTTCCAGAGATAGCACTTTCATGGAATGAC
         GCTATCTTCTAGAATCACTTTTTTTTTTTTTTTGAGTTGGAGTCTCGCTGTGTCGCCAGG
         CTGGAGTGCAGTGGCACAATCTCAGCTCACTGCAATCTCCACCTTCCGGGTTCAAGTGAT
         TCCCCTGCCTCAGCCTCCCGAGGAGCTGTTACTACAGGCGCACACCCCCACTCCTGGCTA
         [-,A]
         TTTTATGTGTTTTAGTAGAGACGGGGTTTCACCGTGTTGGCCAGGATGGTCTCGATCTCC
         TGACTTTGTGATCTGCCTGCTTCAGCCTCCCAAAGTGCTGGGATTACAGGTGTGAGTCAC
         CGCGCCTGGCCTAGAATCACCTTTTTATACCATAACGTGAGCACCACTGCCGCGTCACCA
         AGGAAAGAGAGAGGCAGCTACTGTGGGGTTACAAATGGGTAAGAGTGGCACCAGGAAGGT
         GAAAGTCTCTACTTAGCCAAGGCTTAACAAAATGTCAATCACCAAACATTTATTTATTAA

19670    GACCCCCATGATGAGCAACTATAGCACTAGAACAGTGATAATAACTAATGTTTATAATGC
         ATCTTCAGTTTACAGAGGGCTTTTGTACTCATCATCTAGTTTAGTTCCTGCAACAACCTC
         TTGAGGAATATAGCACAAGCAGGACAAGGGAAGCCCAGAGATGTTAAATAATTTATCCAA
         GTTTATGCTGCTGGGAAGGGCAGCACTGAAATTAAAAGAAAAGTTTTCTGAGCTCAAATC
         CCATGCCCTTTCCTCAATGTGAGCTCTAGCAAGGTATTCAGGAATCCTGCCTCTACAGTT
         [C,T]
         AGAGCCTCAAATTGCTGGGTATGTTGAGTTCTTGTATCTGATTTTTCTAGATTTCCTGCC
         CACATTCTTACTGTCTGGATATCAGGAAAGAGTTTATCAAATGCCTGTGGAAATCCAAGA
         TAAGGTCTCATGATGAGTAACCCAGTGAAAACATGAAGTCAAGTCTAACTAGTCACTACT
         ATTTCACTACTGCTGACTCCTGATGATCAGCTCCTTTTCTAAGTGCTTACTGTCCACTTA
         TTCCATCATCTGCCTAGAATTTATGTGAAGGAATCAAAGCAAAAGGATCATAAGGCTTCC

21153    GGACCCTTGTTTTAGAAGGATGACTGCTGCTATAATGTAGAAAGTGATTTGGAAGAGGGG
         AGGAGTGGGGCACGAAAGATGGTTAGTAGATGGGGGTGGTAATGCTTACCTTTCAGTATT
         TGGAGGCTTCGGAGTCCTCAAAAATTCTCTTCCTTGATTGGAGTCCTCCCAGCCAATAGA
         GGGCTTCACACAAACAGTTTCTTGGGTTTTGAATTGTTTGACCAGAGCTTTCTTCCGACA
         AAAGGTTGGGGTGATTCATTCACTTACCACACCTTGCCTGAACATTCACTTGGGGCTGCC
         [G,T]
         GTTATGAAGGCTATTGTTCTCCAGCCTGTCACAGACGCTTTGAAGACCTGTGCCTCAGCT
         GGTTCTAAGGAGTCAGTTTGTTCAGCTCCGTGCCAGGTTTCCAACTTATGAAATGTGCTG
         GAGATTAACACCTCTCCTGCCATTTTATCCCTACTATAATTGCCAGTCAAAGGATTCCTG
         CAGTTGCCTCTGGCAGCCATAACTGATGAATGTTCTGCCAGCTGCTCTGAGGACCTAGAA
         GAGCAGTTTTCTATCCAGGACCAGTTTCCAAGGGTGGGAGGGTGAAATATATCCTCCAGT

24566    CTACTCTGGAGGCTGAGGTGAGAGGATCACTTGAGTCCAGAAGGTCGAGGTCAAGATTGT
         AGTGAGCCATGATGGCATCACCGCACTCCAGCCTGAGTGACAGAGAGAGACCCTGACTCA
         AAAAAAAAAAAACAAAAAAAAAAAACACCCTCACCACTTATCAGCTATTTGTCTTGAGAA
         TAGTGACATAACCCCTCAGAACCTATTTCCTAATCTGTTAAATGAGGCTGATGACGTTTC
         CTCCTTTTACTGGCAATTTAAACATGATGGATAATAAATGCTAAGCACTTAACACAGGGC
         [C,-]
         TAGAAGATATTAACTGCTCAATAAATGGTAGCTTCTTAACAGTATTCAAACCCATGTGCT
         CTTATCACATGCATTGTTGTCCCTGTGTCCAGTTGGTGGAATGGGAAAAGGCTCCCTTGT
         AACCCCATCTACCATCTTTATCAGACTTTCCTGCCATGGTTCACAGTAAGAGATAGAAGC
         TGCACGGTGACTTCTGGCTCTTTACAATGGTGAGCGGTGTGTGCCTGGTAAGGGAGAGCT
         GATGTCACTGCCCCAAATCCAGTAGTGAGATCTGAGTGTTCTGGTTTCCTCCAGCAGCCT
```

FIG.3-28

```
26604    GATTTGCAGCTGAGCCTGTCTATCTGGTGTGGGAAGAAGATGGGGAGTTACTTGTCAGTC
         CCGGCTTACTTCACCTCCAGAGACCTGTTTCGGTGAGTTGGTCTCCGAGTTCCCCTCTCC
         ATCTCTCCTGGCCCCTGGTCCTGAGAGGAGGGTGGTCTCCCTAAATCTCCTTCTCACTTA
         GTCCTTTACCATCGGTTCTGCCGGGCAGAAGCCAGCGGAGGTTATACCCAAGGAGAATCG
         GCCTTGTGAGGTACCCCCATTATGTCCTGGAAGTGGTGAGGGGAGGGATATACCCAGAAG
         [G,A]
         AACTTCTTAGGGAGCTCCAGCTCCCCTTCTATCCCAGACAAACCTGAAGGAGCCTCCAAA
         AGATGCCACTGACCTGCCCATTGTAGATGTTACTGCTTCCGGGGGGAATAGCCCAAATAG
         AGTGCTGTTTCCAGCTCTCACATGTCTTACCTGCGGGCCATGCTGCCTGCCCAGGAATTT
         GTCCCAACAAGCAGGATGGGCAGGTTTTGCCAAACTGTGGAAACTGGCAAGTCCTGGGTG
         TGGGTAGCCTGGTACACAGTAGGCACCTTATAAACGTTTGTTCTCTTAATGGCAGGCACA

27255    TGGGGAAAGACCTGGGCGAGTGCTTCTAAGACTGGAGCAATGGGCTTTAGAGTGTTCCTG
         AGCTGCTGGGCCAGCCCCCACACCTCCTCAGTCCCTAGGCCTAAGTACCTCCACGAGCCT
         CTCTCTGTGGGGCTTCTCAGAGGGAGATGTGGAAACTCTACCTCTAACCTGGCTTTCTTT
         GCTCATTGCCCCACTCCACCTCCCATAGAAACTCCCCAGGGGGTTTCTGGCCCTCTGGGT
         CCCTTCTGAATGGAGCCATTCCAGGCTAGGGTGGGGTTTGTTTTCATTCTTTGGGAGCAG
         [C,G]
         CTGTTGTTCCAAAAAGGCTGCCTCCCCCTCACCAGTGGTCCTGGTCGACTTTTCCCTTCT
         GGCTTCTCTAAGCTAGGTCCAGTGCCCAGATCTTGCTGCCGGGATACTAGTCAGGTGGCC
         AGGCCCTGGGCAGAAAAGCAGTGTACCATGTGGTTTTGTGGAATGACCGGACCCTGGTAG
         ATTGCTGGGAAGTGTCTGGACAGGGGGAAGGGGGAAGGGAACTGGTCCTCAATGCTGACT
         CTACCAAGCGCCCTGCTAGACACTTTATCCTTTAATCTCTCAACAGCCTAAAGAGATTAT

27399    AGATGTGGAAACTCTACCTCTAACCTGGCTTTCTTTGCTCATTGCCCCACTCCACCTCCC
         ATAGAAACTCCCCAGGGGGTTTCTGGCCCTCTGGGTCCCTTCTGAATGGAGCCATTCCAG
         GCTAGGGTGGGGTTTGTTTTCATTCTTTGGGAGCAGCCTGTTGTTCCAAAAAGGCTGCCT
         CCCCCTCACCAGTGGTCCTGGTCGACTTTTCCCTTCTGGCTTCTCTAAGCTAGGTCCAGT
         GCCCAGATCTTGCTGCCGGGATACTAGTCAGGTGGCCAGGCCCTGGGCAGAAAAGCAGTG
         [T,C]
         ACCATGTGGTTTTGTGGAATGACCGGACCCTGGTAGATTGCTGGGAAGTGTCTGGACAGG
         GGGAAGGGGGAAGGGAACTGGTCCTCAATGCTGACTCTACCAAGCGCCCTGCTAGACACT
         TTATCCTTTAATCTCTCAACAGCCTAAAGAGATTATATATCCCCATTTTACAGATGAGGC
         AACCAGTTTCAACAGAGTTAACATATGGAGCCTCACTGGGCAGCTTTTTCTGTCTTCCTG
         ACTTTCTCTCATCCTTCAGGGGGCTGCAGGTTTGTTTTCTTCTCCTAGTGGAGAGGAAAT

28088    AAGAGCCAATGGAAATTGATCTTGAGTTTAGGAGAAAGCTTTTACATGTGGAATTAAGAT
         GCCAAGTGTTGAAGTAGCCACATTTCAGGTCCTCATTAATTTCTCTTAATCCTGGGAAGG
         CAGCTTAGGAGAAGGGTTGTTCCTTTAGGAGCCAGGAACTATACCCCTTTTACCCTTGGA
         GAGGCAGGGAAGCCAGGGAGGACACAACTTCTCAGGAAGAGGAGAAGCTAGAGCAGATAG
         TGAACTCTCAACCTGAACCTTTAAGGGCCAGACCACTAATGCCACCCAAGTCCACCTGCC
         [G,A]
         TTTGTCTTGTTCTGTCCCAGGCTTTCTGGAGAACCTGATCTTCTTGCCCCTACCCCCAAG
         CTCCGTTTGCCCAGCTAGAGTCTGGGGGGTACTGACTGACTTTCGTAGACATTCTTCCCT
         TCCCCAAATAAGAGGCCACATTCCTGAAGTCACTTCTGAAGAGATAGCTGCCACACAGGG
         CTCTTTCCCCCCAGGGAGGGACCACCCAGACCCTCTGCTCTCCCAGGTATCCGTTACCAC
         ATCACTACCTGGTCAGAAAGCTGTTTCTGCCATTAGCCCCTCCCTCTTTTATTATAGGAT
```

FIG.3-29

```
28734   AAGTAGAAGCTAGACTTCTTGGGCTCCTGAACAGGGTCCTTGCTGGATTCTGTGAAACAA
        ATTAAGTTCTTGACCCTAGGCCTCTGGGGGAGTACAAAGTCTATGGGAGTTCTGGGGCTG
        TGGTTGCAAGGAAAGTGACGCAACCAGATTCCATGGGGACATGATCAGGCGTGACATGTG
        AGGGAGGAAGAGGGAGCAAGGGAATGAAGAATACAACTTCTGTGTCCCATACACCCCTGC
        CTGACAGGCCATACATACTCAGCAGAGAATGCACTGTCTTTCCTACCACACTAGCGTGAG
        [G,A]
        AGTGAGCTGCAATTACCACTGTGCTTCCAAGTAAGAAAATACCTCAAATTGGAATTTACA
        AAAGAGGTAAATTAGGGAGTGGCTTTTGTCGGACATCTTTAAAGCATTTTTCTTTTTATA
        GAATTTCACTTAATGTCCAATACTGATTTAATGAGCTTGGGTTTACACATTATCTCTTGA
        AGAAAACAAATGAACCTTTGTGTTCCAAAGCAATCCATGTTTAAAGGGAAAAAATTATGC
        ATAACTCTGCCCAGCTTCACAGTAACCTTTGGCAGGTGCCTTAGGTCCTCTGGGACTCTT

29246   AATCCATGTTTAAAGGGAAAAAATTATGCATAACTCTGCCCAGCTTCACAGTAACCTTTG
        GCAGGTGCCTTAGGTCCTCTGGGACTCTTTTCCTTATCTGAAAAATGAAGGACTTGGATC
        AGGTGAATGGTTCCCAGCTCTGCAACTTATGTGGCTCCTCAGAGGCACACAAGCTCTTTT
        CCATTATTTGCCAAATAATGGAGGCCCTGTCTTTAACTGCAGTACAACTACACAAAATAC
        TTGAAACTACAGTCTTCCTGGTTTTTGGTTGGAACTGAATCAGTGCACTCTAGCAACACT
        [-,T]
        ATTTCTTGCTGTTCGTAGGCTTCATTATGTGTTTGGTTAATTTTTTAAAACAACAATAAC
        ATATTCCATAATAATTACAGCTTAATTGGCAGACTGTTTCAGTCTATAGGATCTGCAGGA
        AGGAGGAGTAATAAAGGGATTTTTGACTGAGCTCTTATGGAACAGAGTCTCTCTAGGCCC
        CTGTCATATCTGCCCTTCTGGGCCCTGGGGAAAAGTTGGCATCCCCAGTTGTGGTGCTCT
        CCAGGTGCCCTCAGGCTGTGGTGGAGGGAGCTTCCCATTCTCTCCTTCAGCCCACTCAAT

29490   AACTACAGTCTTCCTGGTTTTTGGTTGGAACTGAATCAGTGCACTCTAGCAACACTTATT
        TCTTGCTGTTCGTAGGCTTCATTATGTGTTTGGTTAATTTTTTAAAACAACAATAACATA
        TTCCATAATAATTACAGCTTAATTGGCAGACTGTTTCAGTCTATAGGATCTGCAGGAAGG
        AGGAGTAATAAAGGGATTTTTGACTGAGCTCTTATGGAACAGAGTCTCTCTAGGCCCCTG
        TCATATCTGCCCTTCTGGGCCCTGGGGAAAAGTTGGCATCCCCAGTTGTGGTGCTCTCCA
        [G,A]
        GTGCCCTCAGGCTGTGGTGGAGGGAGCTTCCCATTCTCTCCTTCAGCCCACTCAATTCAG
        AGGCTAGGGGCTGAAAGAAGCTTCTCTACAACTGGCTGTTCACTGGGAGGTTAAGGGATG
        ACCATCCAGCCAGGCCTTCCTCAGGACATGGGAGGGCTTATGCTTTAACATGTGTAAATC
        CACTGCAATAATGACTGGTTCTTTTACCCCATAAGGTTGAGAATTTACCTGTAAACATTT
        TTGTCTGAAGAATTTGGATGTAAGTGAGGGCTGGGCCTCTATCTTATCTCACTTGGCTTC

29934   GGACATGGGAGGGCTTATGCTTTAACATGTGTAAATCCACTGCAATAATGACTGGTTCTT
        TTACCCCATAAGGTTGAGAATTTACCTGTAAACATTTTTGTCTGAAGAATTTGGATGTAA
        GTGAGGGCTGGGCCTCTATCTTATCTCACTTGGCTTCTCTCAGCACAGCACCTTGCCTGC
        TTGTTCTTACACATCCTAGATGCACAGTAACTATTTCCTAATTATTAGAAATCTATTAGA
        ATCAATTGATTTCAGCTGGGCTTGGTGGCTCCTTCCTGTAATCCCAGCACTTTGGGAGGC
        [T,C]
        AAGGCTGGAGGATCACCTGAGTCCAGGAGTTTAAGACCAGCCTGGGCAACATAGGGAGAC
        CCTGTCTCTACAAAAAATAAAAAATTAGCCAGGCATGGTGGTGTGCACCTGTAGTCCCAG
        CTACTCAGGAGGCTGAGGCAGGAGGATCTCTTGAGCCTGGGAGGTCAGACTACAGTGAGC
        AATGATTGTGCCACTGCACTCCAGCCTGGGTGACAGAGTAAGACTCTGTCTCTTAAAAAA
        AAAAAAAAAAAAGTTGATTTCTATTTGGATAGATAAATAATTCATTTTAGGACCTTTCTT
```

FIG.3-30

```
34480   CTGACTTCAAGTGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTATAAGCATAAGC
        CACTGTGCCCAGCTGCTCTCTATATTTTTAATACATATTATTTCCATTAATTTTCACAGC
        AGTTCATTTTATAGATGAGGAAACTAGGCCAGAGAAGTAAAATATCTTGCCCAAGATGAT
        GTAACTAGTAAGTGGCAGGATCAAGATTCAAACCAAGCAATGTTCAAACCTCTTGGAAGC
        AAGAATGTGGCCACTGTGGAAGGTGCAAGGCCTTGACAACAAGAATAGGGAAAAGAAGGA
        [A,G]
        CTAGAAGGAAAGAGATGGCATGGGCTCAGCAGGCCAGGGAGCTCTTAGCTGTGTGTGTTG
        GGAAGCTCAGAAGGGAGGAAGAGGTTGTCTGTGCAGGTAAGTCCTGAGAACACACCAGAC
        TTTTGAGAGGTGGAGCTTCATAGCCAGGTCATTAGGGGAGAAGGGAGCTATAGATTTTTT
        TTTTTTTTTTTTTTTTTTTTTTTTTTTTAGAGACGGGGTCTTACTATGTTGCCCAGGCTG
        GTCTTGAACTCCTGGGCTCAAGTGATCCTCCCACCTCAGCCTCCCAAAGTGCTGGGATTA

38812   AAATCCAGCAGATCCATTGAGAGTTTAAGCAGCAAGGTGTTGTGACCAAGTTAACATTTT
        AGAAGGATCACTGGTATGGAGGTTGGATTGGAGAGGGGAAAGCCTAAAGGTATAGAGACT
        AGTTAGGAAGCTATTGTAGGCTGGGCATGGTGGTTCATGCCTGTAATCTCAGCACTTTGG
        GAGGCTGAGGTGGGAGGATTGCTTGAGGCCAGGAGTTGAAGACCAACCTGGCCAACATAG
        CAAGACCCCGTCTCTGTTTTTCTTAATTAAAAGAAAAGTCCAGACGTAGACATAGTGGCT
        [T,C]
        ACGCCTGTAATGCCAGCACTTTGGGAGGCCAAGGTGGGCAGATTGCTTGAGGTCAAGAGT
        TTGGGATTAGGCCAGGCGCAGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAG
        GTGGGCGGATCACAAGGTCAGGAGATCAAGACCATCCTGGCTAACACAATGAAACCCCGT
        CTCTACTAAAAGTACAAAAATTAGCCGGGCATGGTGGCGGACGCCTGTAGTCCCAGCTAC
        TCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCTAGGAGGCGGAGCTTGCTGTGAGCAGA

40731   GTTCTGTCCTATGTCTGTCTCTCGGATGAAGCTGAGCTGGCTTTCAGAAGCCTGCAGAGT
        TAGGAAAGGAACCAGCTGGCCAGGGACAGACTATGAGGATTGTGCTGACCCAGCTGCCCC
        TGTGGGGATCACAGTTTACAGCCAGAGCCTGTGCGGACCCAGCTGTCTGCCAGGTTTCCT
        TAGAAACCTGAGAGTCAGTCTCTGTCCACTGAACTCCTAAGCTGGACAGGAGGCAGTGAT
        GCTAAACCCTGAAGGGCAACATGGCCTATGGAGAAAGCATGGAGCTCAGAGCCTGGAGTA
        [C,G]
        GGGCACAGATAGGATTGAATAAATTGTGTAGAAAGACTTTGAAAACAATAAAGCAAAAGA
        TGAATGAACGTTTTTTTTAGACTTGAGGGACCAACAACCCCCAAACCCCAGATTCTGCCA
        GGTCCATGGGGAAGGAGAAGTTGCCTTGAGTGGAAGCCCCAAGTAGGGAGACTTACAGAA
        AAGAAGTCAAGAGCACTGGCTCCCAGGCAGAAATACTGATACCCTACTGGGGCTTCAGGC
        TGAGCTCCTCCCTTCACAAATCACTTCATCTCTCTGAGCCTGTTTCTGCATCTGTGACAT

41303   CTCTGAGCCTGTTTCTGCATCTGTGACATAAGATGGTAAGATAAAGGTGGCTGTCTCACC
        AATTATGTAAGGATTAAATGTGGAAAAGGACATAAAGTTGTATAGTGCTGCCATAGGGAC
        AGTGTTCAGTAAACGTGACACATTCTTAGTATCACTAAGAATCAGGTTCTTGGCCAGGCA
        CCGTGGCTCATGCCTGTAATCCCAACACTCTGGGAGGCCTAGGTCGGAGGATGGCTTGAA
        CACAGGAGTTTGAGACCAGCCTGAGCAACATAGTGAGACACTGTCTCTACAAAAAAAAAA
        [T,A]
        AATAATAATAATTGTTTTTAATTAGATGGGCAGGGCACTGTGGCTCACACCTGTAATCCC
        AGCACTTTGGGAGGCCAAGGCCGGAGGATTGCTTGAGGCCAGGAGTTCAGGAGCAGCCTG
        GGCCACATTCCTGTCTCTACAAAGAATAAAAAAGTTAACTGGGCATGGTGGCACATGCCT
        GTAATCCCAGCTACTCAAGAGGCTGAGGAGGAGGATTGCCTGAGCCCAGGAGTTCAAGAC
        TGCAGTGAGCCTTGATCACACCACTGTACTACAGCTTGGGCAACAGAGTGAGACCTTGTC
```

FIG.3-31

```
41305     CTGAGCCTGTTTCTGCATCTGTGACATAAGATGGTAAGATAAAGGTGGCTGTCTCACCAA
          TTATGTAAGGATTAAATGTGGAAAAGGACATAAAGTTGTATAGTGCTGCCATAGGGACAG
          TGTTCAGTAAACGTGACACATTCTTAGTATCACTAAGAATCAGGTTCTTGGCCAGGCACC
          GTGGCTCATGCCTGTAATCCCAACACTCTGGGAGGCCTAGGTCGGAGGATGGCTTGAACA
          CAGGAGTTTGAGACCAGCCTGAGCAACATAGTGAGACACTGTCTCTACAAAAAAAAAATA
          [-,A]
          TAATAATAATTGTTTTTAATTAGATGGGCAGGGCACTGTGGCTCACACCTGTAATCCCAG
          CACTTTGGGAGGCCAAGGCCGGAGGATTGCTTGAGGCCAGGAGTTCAGGAGCAGCCTGGG
          CCACATTCCTGTCTCTACAAAGAATAAAAAAGTTAACTGGGCATGGTGGCACATGCCTGT
          AATCCCAGCTACTCAAGAGGCTGAGGAGGAGGATTGCCTGAGCCCAGGAGTTCAAGACTG
          CAGTGAGCCTTGATCACACCACTGTACTACAGCTTGGGCAACAGAGTGAGACCTTGTCTC

41457     CTAAGAATCAGGTTCTTGGCCAGGCACCGTGGCTCATGCCTGTAATCCCAACACTCTGGG
          AGGCCTAGGTCGGAGGATGGCTTGAACACAGGAGTTTGAGACCAGCCTGAGCAACATAGT
          GAGACACTGTCTCTACAAAAAAAAAATAATAATAATAATTGTTTTTAATTAGATGGGCAG
          GGCACTGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCAAGGCCGGAGGATTGCT
          TGAGGCCAGGAGTTCAGGAGCAGCCTGGGCCACATTCCTGTCTCTACAAAGAATAAAAAA
          [G,C]
          TTAACTGGGCATGGTGGCACATGCCTGTAATCCCAGCTACTCAAGAGGCTGAGGAGGAGG
          ATTGCCTGAGCCCAGGAGTTCAAGACTGCAGTGAGCCTTGATCACACCACTGTACTACAG
          CTTGGGCAACAGAGTGAGACCTTGTCTCCAAAAAAAAAAGTTTGTTTTTTTTTATCCACT
          CTCCTCACCAAACAAACTGAGTAAGTTAGAGCCCTCTCAGCTGGCATGTGTTGGAAACAG
          TGCCCTCTCATTAAAGTGCTGCCCTCACTCCCATTGCCTCTTGGCCTTGGTCAGTATGAT

43168     AGCTACTTGGGAGGCTGAGGCAGGAGAATCGCTTGAACCTGGAAGGCGGAGGTCGCAGTG
          AGCCGAGATCGTGCCATTGCACTTCAGCCTGGGCGACAGAGCGAGACTCTGTCTCAAAAA
          TAATAATAATAACAATAACTAGCCGGGCCTGGTGGCACATGCCTGTAGTCCCAGTTACTC
          AGGAGGCGGAGGCATGAGACTCAGGTGAACTAGGGAGACAGAGGTTGCAGTGAGCCAAGA
          TCACACCACTGCACTCCAGCCTGGTTGACAGAGCGAGACTCTGTCTCAAAAAAAAAAAAA
          [A,-,T]
          CCCATTTGCTCATTTTTTGGATACTAGTATAACTATCACTCTAAACCAGTTAGTACTTAA
          ATCAAGCAGATATGGGAGATGGTGAATTACCATCTACAGTGTTGTCATATATGTCACATA
          CTGAGCATTATCAGCTAGTAGAATCTAGTTAATTGTTCTATGTGTGATGTATGCAGAGTT
          CCCATTTTGAATGTGTTTTTACTATGCTTAAATAAATGACTGATGTCAGCAACCCCAAAA
          TGATACATCTGATGTAAGAGCCCCTGTTCCCCAATAATAACATCTAAACTATAGACATTG

43357     AGGCATGAGACTCAGGTGAACTAGGGAGACAGAGGTTGCAGTGAGCCAAGATCACACCAC
          TGCACTCCAGCCTGGTTGACAGAGCGAGACTCTGTCTCAAAAAAAAAAAAATCCCATTTG
          CTCATTTTTTGGATACTAGTATAACTATCACTCTAAACCAGTTAGTACTTAAATCAAGCA
          GATATGGGAGATGGTGAATTACCATCTACAGTGTTGTCATATATGTCACATACTGAGCAT
          TATCAGCTAGTAGAATCTAGTTAATTGTTCTATGTGTGATGTATGCAGAGTTCCCATTTT
          [T,G]
          AATGTGTTTTTACTATGCTTAAATAAATGACTGATGTCAGCAACCCCAAAATGATACATC
          TGATGTAAGAGCCCCTGTTCCCCAATAATAACATCTAAACTATAGACATTGGAATGAACA
          GGTGCCCCTAAGTTTCCTCCCTCCAGGGTTTCTTGGCCGGTCTCTGAGGACTACACATCC
          CTACTCCCGTCTTTCCTCATCTTCAGGCGCAGTAACAGTATCTCCAAGTCCCCTGGCCCC
          AGCTCCCCAAAGGAGCCCCTGCTGTTCAGCCGTGACATCAGCCGCTCAGAATCCCTTCGT
```

FIG.3-32

```
45664   CCAGCTTTCCTTGGCTTCCCCCACCCCCAGGTGAAAGTGATGCGCAGCCTGGACCACCCC
        AATGTGCTCAAGTTCATTGGTGTGCTGTACAAGGATAAGAAGCTGAACCTGCTGACAGAG
        TACATTGAGGGGGGCACACTGAAGGACTTTCTGCGCAGTATGGTGAGCACACCACCCCAT
        AGTCTCCAGGAGCCTTGGTGGGTTGTCAGACACCTATGCTATCACTACCCTAGGAGCTTA
        AAGGGCAGAGGGGCCCTGCTTTGCCTCCAAAGGACCATGCTGGGTGGGACTGAGCATACA
        [T,C]
        AGGGAGGCTTCACTGGGAGACCACATTGACCCATGGGGCCTGGACCACGAGTGGGACAGG
        GCTCAACAGCCTCTGAAAATCATTCCCCATTCTGCAGGATCCGTTCCCCTGGCAGCAGAA
        GGTCAGGTTTGCCAAAGGAATCGCCTCCGGAATGGTGAGTCCCACCAACAAACCTGCCAG
        CAGGGCGAGAGTAGGGAGAGGTGTGAGAATTGTGGGCTTCACTGGAAGGTAGAGACCCCT
        TCCTATGCAACTTGTGTGGGCTGGGTCAGCAGCTATTCATTGAGTTTGTCTGTGTCACTG

47549   AATTAGCTGGGCGTGGTGGTGCACGCCTGTAGTCCCAGCTACTCAGGAGGCCGAGGCAGG
        AGAATAGCTTGAACCTGGGAGGCAGAAGTTGCAGTGAGCCAAGATCACACCACTGCATTC
        CAGCCTGGGTGACAGAGTGAGACTTCATCTCAAAAAAAAAAAAAAAGAGAGACTGATATG
        GTTAGTACATTGGGGTGGAATGCGGAGGGTCCAGGGAATGGAGCCCTGCATAGGGGGCTA
        ATGAAACATTTCAGATTTCTGAATTAAGGTAGTGGCTGTGGGGACAGGAGCCTGGGAGGC
        [A,C]
        GGGTGGAGTCAGAATGGAGAGACTGGTTGGCAATGAGGGAACAGGAGGAGGAGGAGGAGG
        AGTTACGAGTGGCTTGAGGTGTCACTTACCAGACATTTGGGGGATGGGGGATAGCCGTGA
        TTGTTGAGCAACTGGTTTGGGAAGAGCTAGCATTGATCCCTGCTGTTCTGTGCTAGCAGA
        ACCTATCAGCATCTTCTGGGCAGGAAACTGGCTCCATGAGACTGGCTTAGGGAGAGGCTG
        CTAGTCACCTAATCTGCAGAGAAGGGGCAGCTGGAGCTGTGGGACAGAAGAGGCATCCAT

47908   GGAGTTACGAGTGGCTTGAGGTGTCACTTACCAGACATTTGGGGGATGGGGGATAGCCGT
        GATTGTTGAGCAACTGGTTTGGGAAGAGCTAGCATTGATCCCTGCTGTTCTGTGCTAGCA
        GAACCTATCAGCATCTTCTGGGCAGGAAACTGGCTCCATGAGACTGGCTTAGGGAGAGGC
        TGCTAGTCACCTAATCTGCAGAGAAGGGGCAGCTGGAGCTGTGGGACAGAAGAGGCATCC
        ATGTAGCTGGTGGGGGTGTCTCAGCTTGTGAAGAGGAGATGGCTTTGAGCAGGGCTGACA
        [C,A]
        TGAAAAGGCTGGAAGAAAAAAACAGACACACAAGAGTCTCAGGATCAGGTAGCATAGGAA
        AGTTGTGGACAGTCTTTGAGGAGCACTCCCTCAGGCAGGCAGGCAGGCAGGTCATGAGCT
        ATAGCGATTCAGGAAGAGCTCCCTGGGTGTGTGAGCAGCTCCAGGAGCCTAAGGGATGAA
        AGTAGTATTGCAGGGGGCTGGAGAGCAAGGAGTGGCTCCTTCTACATTTGCAAGGGAAGG
        AGAAAGGAAGTTGCTCCTGAGAGTGGTAAGAGTCAGTGGTGGAGGCCTGGAGAGGAGACA

52267   TTGTGAGGGGTAGAGGAGAGGAGAGACAAGGGATGGTTAGGATAATGAAGGAATGTTTTG
        TTTTTGTTTTTGTTTTTGAGATGGAGTTTCACTCTGTCACCCAGGCTGGAGTGCAGAGGT
        GCAATCTTGGCTCACTGCAGCCTCCGCCTCCCAGGTTCAAGCAATCCTCCTGCCTCAGCC
        TCCCAAGTAGCTGGGACTACAGGTGTGCGCCACCACGCCTGGCTAATTTTTGTATTTTCA
        GTAGAGACAGGGTTTCGCCATATTGGCCAGGCTGGTCTCAAATGCCTGACCTCAGGTGAT
        [C,A]
        CACCCGCTTCAGCCTCCCAAAGTGCTGAGATTACAGGCATGAGCTACCGTGCCTGGCCAT
        GAAGGAAGATTTGTTTTAAAAAATTGTTTTCTTTAATATTAATTGAACACCTCTGTTCAG
        AGCACTGGGCTGGTGCCAGAGGGTTTCAGACATGAATCAGATCCAGCACCTCATAGAGCC
        TTAATCTGGCACACACACACAGCCACAAGGAGACACAGACAAGGCAGGGTAGGATGAGTG
        GAAGCTAGGAGCAGATGCTGATTTGGAACACTTGGCTTCTGCAGTGAAGCCCCTTCTTAG
```

FIG.3-33

```
54654     GGCCCCGGCCCCGGCCCCCAGGCCAGGCAGTGGCGGCCAAGGACCACGCATCTACTTTCA
          GAGCCCCCCCGGGGCCGCAGGAGAGGGCCCGGGCTGGGCGGATGATGAGGGCCCAGTGA
          GGCGCCAAGGGAAGGTCACCATCAAGTATGACCCCAAGGAGCTACGGAAGCACCTCAACC
          TAGAGGAGTGGATCCTGGAGCAGCTCACGCGCCTCTACGACTGCCAGGAAGAGGAGATCT
          CAGAACTAGAGATTGACGTGGATGAGCTCCTGGACATGGAGAGTGACGATGCCTGGGCTT
          [T,C]
          CAGGGTCAAGGAGCTGCTGGTTGACTGTTACAAACCCACAGAGGCCTTCATCTCTGGCCT
          GCTGGACAAGATCCGGGCCATGCAGAAGCTGAGCACACCCCAGAAGAAGTGAGGGTCCCC
          GACCCAGGCGAACGGTGGCTCCCATAGGACAATCGCTACCCCCCGACCTCGTAGCAACAG
          CAATACCGGGGGACCCTGCGGCCAGGCCTGGTTCCATGAGCAGGGCTCCTCGTGCCCCTG
          GCCCAGGGGTCTCTTCCCCTGCCCCCTCAGTTTTCCACTTTTGGATTTTTTTATTGTTAT

54679     GGCAGTGGCGGCCAAGGACCACGCATCTACTTTCAGAGCCCCCCCGGGGCCGCAGGAGA
          GGGCCCGGGCTGGGCGGATGATGAGGGCCCAGTGAGGCGCCAAGGGAAGGTCACCATCAA
          GTATGACCCCAAGGAGCTACGGAAGCACCTCAACCTAGAGGAGTGGATCCTGGAGCAGCT
          CACGCGCCTCTACGACTGCCAGGAAGAGGAGATCTCAGAACTAGAGATTGACGTGGATGA
          GCTCCTGGACATGGAGAGTGACGATGCCTGGGCTTCCAGGGTCAAGGAGCTGCTGGTTGA
          [C,G]
          TGTTACAAACCCACAGAGGCCTTCATCTCTGGCCTGCTGGACAAGATCCGGGCCATGCAG
          AAGCTGAGCACACCCCAGAAGAAGTGAGGGTCCCCGACCCAGGCGAACGGTGGCTCCCAT
          AGGACAATCGCTACCCCCCGACCTCGTAGCAACAGCAATACCGGGGGACCCTGCGGCCAG
          GCCTGGTTCCATGAGCAGGGCTCCTCGTGCCCCTGGCCCAGGGGTCTCTTCCCCTGCCCC
          CTCAGTTTTCCACTTTTGGATTTTTTTATTGTTATTAAACTGATGGGACTTTGTGTTTTT

54693     AGGACCACGCATCTACTTTCAGAGCCCCCCCGGGGCCGCAGGAGAGGGCCCGGGCTGGG
          CGGATGATGAGGGCCCAGTGAGGCGCCAAGGGAAGGTCACCATCAAGTATGACCCCAAGG
          AGCTACGGAAGCACCTCAACCTAGAGGAGTGGATCCTGGAGCAGCTCACGCGCCTCTACG
          ACTGCCAGGAAGAGGAGATCTCAGAACTAGAGATTGACGTGGATGAGCTCCTGGACATGG
          AGAGTGACGATGCCTGGGCTTCCAGGGTCAAGGAGCTGCTGGTTGACTGTTACAAACCCA
          [A,C]
          AGAGGCCTTCATCTCTGGCCTGCTGGACAAGATCCGGGCCATGCAGAAGCTGAGCACACC
          CCAGAAGAAGTGAGGGTCCCCGACCCAGGCGAACGGTGGCTCCCATAGGACAATCGCTAC
          CCCCCGACCTCGTAGCAACAGCAATACCGGGGGACCCTGCGGCCAGGCCTGGTTCCATGA
          GCAGGGCTCCTCGTGCCCCTGGCCCAGGGGTCTCTTCCCCTGCCCCCTCAGTTTTCCACT
          TTTGGATTTTTTTATTGTTATTAAACTGATGGGACTTTGTGTTTTTATATTGACTCTGCG

54706     TACTTTCAGAGCCCCCCCGGGGCCGCAGGAGAGGGCCCGGGCTGGGCGGATGATGAGGG
          CCCAGTGAGGCGCCAAGGGAAGGTCACCATCAAGTATGACCCCAAGGAGCTACGGAAGCA
          CCTCAACCTAGAGGAGTGGATCCTGGAGCAGCTCACGCGCCTCTACGACTGCCAGGAAGA
          GGAGATCTCAGAACTAGAGATTGACGTGGATGAGCTCCTGGACATGGAGAGTGACGATGC
          CTGGGCTTCCAGGGTCAAGGAGCTGCTGGTTGACTGTTACAAACCCACAGAGGCCTTCAT
          [T,C]
          TCTGGCCTGCTGGACAAGATCCGGGCCATGCAGAAGCTGAGCACACCCCAGAAGAAGTGA
          GGGTCCCCGACCCAGGCGAACGGTGGCTCCCATAGGACAATCGCTACCCCCCGACCTCGT
          AGCAACAGCAATACCGGGGGACCCTGCGGCCAGGCCTGGTTCCATGAGCAGGGCTCCTCG
          TGCCCCTGGCCCAGGGGTCTCTTCCCCTGCCCCCTCAGTTTTCCACTTTTGGATTTTTTT
          ATTGTTATTAAACTGATGGGACTTTGTGTTTTTATATTGACTCTGCGGCACGGGCCCTTT
```

FIG.3-34

```
54712   CAGAGCCCCCCCGGGGCCGCAGGAGAGGGCCCGGGCTGGGCGGATGATGAGGGCCCAGT
        GAGGCGCCAAGGGAAGGTCACCATCAAGTATGACCCCAAGGAGCTACGGAAGCACCTCAA
        CCTAGAGGAGTGGATCCTGGAGCAGCTCACGCGCCTCTACGACTGCCAGGAAGAGGAGAT
        CTCAGAACTAGAGATTGACGTGGATGAGCTCCTGGACATGGAGAGTGACGATGCCTGGGC
        TTCCAGGGTCAAGGAGCTGCTGGTTGACTGTTACAAACCCACAGAGGCCTTCATCTCTGG
        [T,C]
        CTGCTGGACAAGATCCGGGCCATGCAGAAGCTGAGCACACCCCAGAAGAAGTGAGGGTCC
        CCGACCCAGGCGAACGGTGGCTCCCATAGGACAATCGCTACCCCCCGACCTCGTAGCAAC
        AGCAATACCGGGGGACCCTGCGGCCAGGCCTGGTTCCATGAGCAGGGCTCCTCGTGCCCC
        TGGCCCAGGGGTCTCTTCCCCTGCCCCCTCAGTTTTCCACTTTTGGATTTTTTTATTGTT
        ATTAAACTGATGGGACTTTGTGTTTTTATATTGACTCTGCGGCACGGGCCCTTTAATAAA

54799   GTATGACCCCAAGGAGCTACGGAAGCACCTCAACCTAGAGGAGTGGATCCTGGAGCAGCT
        CACGCGCCTCTACGACTGCCAGGAAGAGGAGATCTCAGAACTAGAGATTGACGTGGATGA
        GCTCCTGGACATGGAGAGTGACGATGCCTGGGCTTCCAGGGTCAAGGAGCTGCTGGTTGA
        CTGTTACAAACCCACAGAGGCCTTCATCTCTGGCCTGCTGGACAAGATCCGGGCCATGCA
        GAAGCTGAGCACACCCCAGAAGAAGTGAGGGTCCCCGACCCAGGCGAACGGTGGCTCCCA
        [T.C]
        AGGACAATCGCTACCCCCCGACCTCGTAGCAACAGCAATACCGGGGGACCCTGCGGCCAG
        GCCTGGTTCCATGAGCAGGGCTCCTCGTGCCCCTGGCCCAGGGGTCTCTTCCCCTGCCCC
        CTCAGTTTTCCACTTTTGGATTTTTTTATTGTTATTAAACTGATGGGACTTTGTGTTTTT
        ATATTGACTCTGCGGCACGGGCCCTTTAATAAAGCGAGGTAGGGTACGCCTTTGGTGCAG
        CTCAAAAAAAAAAAAAAAAATGATTTCCAGCGGTCCACATTAGAGTTGAAATTTTCTGGT

54819   GGAAGCACCTCAACCTAGAGGAGTGGATCCTGGAGCAGCTCACGCGCCTCTACGACTGCC
        AGGAAGAGGAGATCTCAGAACTAGAGATTGACGTGGATGAGCTCCTGGACATGGAGAGTG
        ACGATGCCTGGGCTTCCAGGGTCAAGGAGCTGCTGGTTGACTGTTACAAACCCACAGAGG
        CCTTCATCTCTGGCCTGCTGGACAAGATCCGGGCCATGCAGAAGCTGAGCACACCCCAGA
        AGAAGTGAGGGTCCCCGACCCAGGCGAACGGTGGCTCCCATAGGACAATCGCTACCCCCC
        [G,A]
        ACCTCGTAGCAACAGCAATACCGGGGGACCCTGCGGCCAGGCCTGGTTCCATGAGCAGGG
        CTCCTCGTGCCCCTGGCCCAGGGGTCTCTTCCCCTGCCCCCTCAGTTTTCCACTTTTGGA
        TTTTTTTATTGTTATTAAACTGATGGGACTTTGTGTTTTTATATTGACTCTGCGGCACGG
        GCCCTTTAATAAAGCGAGGTAGGGTACGCCTTTGGTGCAGCTCAAAAAAAAAAAAAAAAA
        TGATTTCCAGCGGTCCACATTAGAGTTGAAATTTTCTGGTGGGAGAATCTATACCTTGTT

55499   TTGTTTTCTAATACCTCTTGTCATTCTAAATATCTTTAATTTATTAAAAAATATATATAT
        ACAGTATTGAATGCCTACTGTGTGCTAGGTACAGTTCTAAACACTTGGGTTACAGCAGCG
        AACAAAATAAAGGTGCTTACCCTCATAGAACATAGATTCTAGCATGGTATCTACTGTATC
        ATACAGTAGATACAATAAGTAAACTATATTGAATATTAGAATGTGGCAGATGCTATGGAA
        AAAGAGTCAAGACAAGTAAAGACGATTGTTCAGGGTACCAGTTGCAATTTTAAATATGGT
        [C,T]
        GTCAGAGCAGGCCTCACTGAGGTGACATGACATTTAAGCATAAACATGGAGGAGGAGGAG
        TAAGCCTGAGCTGTCTTAGGCTTCCGGGGCAGCCAAGCCATTTCCGTGGCACTAGGAGCC
        TGGTGTTTCCGATTCCACCTTTGATAACTGCATTTTCTCTAAGATATGGGAGGGAAGTTT
        TTCTCCTATTGTTTTTAAGTATTAACTCCAGCTAGTCCAGCCTTGTTATAGTGTTACCTA
        ATCTTTATAGCAAATATATGAGGTACCGGTAACATTATGCCCATTTCTCACAGAGGCACT
```

FIG.3-35

```
56825   ACTGATGGCTCAAAGGGTGTGAAAAAGTCAGTGATGCTCCCCCTTTCTACTCCAGATCCT
        GTCCTTCCTGGAGCAAGGTTGAGGGAGTAGGTTTTGAAGAGTCCCTTAATATGTGGTGGA
        ACAGGCCAGGAGTTAGAGAAAGGGCTGGCTTCTGTTTACCTGCTCACTGGCTCTAGCCAG
        CCCAGGGACCACATCAATGTGAGAGGAAGCCTCCACCTCATGTTTTCAAACTTAATACTG
        GAGACTGGCTGAGAACTTACGGACAACATCCTTTCTGTCTGAAACAAACAGTCACAAGCA
        [C,A]
        AGGAAGAGGCTGGGGGACTAGAAAGAGGCCCTGCCCTCTAGAAAGCTCAGATCTTGGCTT
        CTGTTACTCATACTCGGGTGGGCTCCTTAGTCAGATGCCTAAAACATTTTGCCTAAAGCT
        CGATGGGTTCTGGAGGACAGTGTGGCTTGTCACAGGCCTAGAGTCTGAGGGAGGGGAGTG
        GGAGTCTCAGCAATCTCTTGGTCTTGGCTTCATGGCAACCACTGCTCACCCTTCAACATG
        CCTGGTTTAGGCAGCAGCTTGGGCTGGGAAGAGGTGGTGGCAGAGTCTCAAAGCTGAGAT

58871   CGTCACCCACCACCCAACCCCTGCCGCACTCCAGCCTTTAACAAGGGCTGTCTAGATATT
        CATTTTAACTACCTCCACCTTGGAAACAATTGCTGAAGGGGAGAGGATTTGCAATGACCA
        ACCACCTTGTTGGGACGCCTGCACACCTGTCTTTCCTGCTTCAACCTGAAAGATTCCTGA
        TGATGATAATCTGGACACAGAAGCCGGGCACGGTGGCTCTAGCCTGTAATCTCAGCACTT
        TGGGAGGCCTCAGCAGGTGGATCACCTGAGATCAAGAGTTTGAGAACAGCCTGACCAACA
        [T,A]
        GGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCAGGTGTGGTGGCACATACCTG
        TAATCCCAGCTACTCTGGAGGCTGAGGCAGGAGAATCGCTTGAACCCACAAGGCAGAGGT
        TGCAGTGAGGCGAGATCATGCCATTGCACTCCAGCCTGTGCAACAAGAGCCAAACTCCAT
        CTCAAAAAAAAAAA
```

FIG.3-36

ISOLATED HUMAN KINASE PROTEINS, NUCLEIC ACID MOLECULES ENCODING HUMAN KINASE PROTEINS, AND USES THEREOF

FIELD OF THE INVENTION

The present invention is in the field of kinase proteins that are related to the serine/threonine kinase subfamily, recombinant DNA molecules, and protein production. The present invention specifically provides novel peptides and proteins that effect protein phosphorylation and nucleic acid molecules encoding such peptide and protein molecules, all of which are useful in the development of human therapeutics and diagnostic compositions and methods.

BACKGROUND OF THE INVENTION

Protein Kinases

Kinases regulate many different cell proliferation, differentiation, and signaling processes by adding phosphate groups to proteins. Uncontrolled signaling has been implicated in a variety of disease conditions including inflammation, cancer, arteriosclerosis, and psoriasis. Reversible protein phosphorylation is the main strategy for controlling activities of eukaryotic cells. It is estimated that more than 1000 of the 10,000 proteins active in a typical mammalian cell are phosphorylated. The high energy phosphate, which drives activation, is generally transferred from adenosine triphosphate molecules (ATP) to a particular protein by protein kinases and removed from that protein by protein phosphatases. Phosphorylation occurs in response to extracellular signals (hormones, neurotransmitters, growth and differentiation factors, etc), cell cycle checkpoints, and environmental or nutritional stresses and is roughly analogous to turning on a molecular switch. When the switch goes on, the appropriate protein kinase activates a metabolic enzyme, regulatory protein, receptor, cytoskeletal protein, ion channel or pump, or transcription factor.

The kinases comprise the largest known protein group, a superfamily of enzymes with widely varied functions and specificities. They are usually named after their substrate, their regulatory molecules, or some aspect of a mutant phenotype. With regard to substrates, the protein kinases may be roughly divided into two groups; those that phosphorylate tyrosine residues (protein tyrosine kinases, PTK) and those that phosphorylate serine or threonine residues (serine/threonine kinases, STK). A few protein kinases have dual specificity and phosphorylate threonine and tyrosine residues. Almost all kinases contain a similar 250–300 amino acid catalytic domain. The N-terminal domain, which contains subdomains I–IV, generally folds into a two-lobed structure, which binds and orients the ATP (or GTP) donor molecule. The larger C terminal lobe, which contains subdomains VI A–XI, binds the protein substrate and carries out the transfer of the gamma phosphate from ATP to the hydroxyl group of a serine, threonine, or tyrosine residue. Subdomain V spans the two lobes.

The kinases may be categorized into families by the different amino acid sequences (generally between 5 and 100 residues) located on either side of, or inserted into loops of, the kinase domain. These added amino acid sequences allow the regulation of each kinase as it recognizes and interacts with its target protein. The primary structure of the kinase domains is conserved and can be further subdivided into 11 subdomains. Each of the 11 subdomains contains specific residues and motifs or patterns of amino acids that are characteristic of that subdomain and are highly conserved (Hardie, G. and Hanks, S. (1995) *The Protein Kinase Facts Books*, Vol I:7–20 Academic Press, San Diego, Calif.).

The second messenger dependent protein kinases primarily mediate the effects of second messengers such as cyclic AMP (cAMP), cyclic GMP, inositol triphosphate, phosphatidylinositol, 3,4,5-triphosphate, cyclic-ADPribose, arachidonic acid, diacylglycerol and calcium-calmodulin. The cyclic-AMP dependent protein kinases (PKA) are important members of the STK family. Cyclic-AMP is an intracellular mediator of hormone action in all prokaryotic and animal cells that have been studied. Such hormone-induced cellular responses include thyroid hormone secretion, cortisol secretion, progesterone secretion, glycogen breakdown, bone resorption, and regulation of heart rate and force of heart muscle contraction. PKA is found in all animal cells and is thought to account for the effects of cyclic-AMP in most of these cells. Altered PKA expression is implicated in a variety of disorders and diseases including cancer, thyroid disorders, diabetes, atherosclerosis, and cardiovascular disease (Isselbacher, K. J. et al. (1994) *Harrison's Principles of Internal Medicine*, McGraw-Hill, New York, N.Y., pp. 416–431, 1887).

Calcium-calmodulin (CaM) dependent protein kinases are also members of STK family. Calmodulin is a calcium receptor that mediates many calcium regulated processes by binding to target proteins in response to the binding of calcium. The principle target protein in these processes is CaM dependent protein kinases. CaM-kinases are involved in regulation of smooth muscle contraction (MLC kinase), glycogen breakdown (phosphorylase kinase), and neurotransmission (CaM kinase I and CaM kinase II). CaM kinase I phosphorylates a variety of substrates including the neurotransmitter related proteins synapsin I and II, the gene transcription regulator, CREB, and the cystic fibrosis conductance regulator protein, CFTR (Haribabu, B. et al. (1995) *EMBO Journal* 14:3679–86). CaM II kinase also phosphorylates synapsin at different sites, and controls the synthesis of catecholamines in the brain through phosphorylation and activation of tyrosine hydroxylase. Many of the CaM kinases are activated by phosphorylation in addition to binding to CaM. The kinase may autophosphorylate itself, or be phosphorylated by another kinase as part of a "kinase cascade".

Another ligand-activated protein kinase is 5'-AMP-activated protein kinase (AMPK) (Gao, G. et al. (1996) *J. Biol Chem.* 15:8675–81). Mammalian AMPK is a regulator of fatty acid and sterol synthesis through phosphorylation of the enzymes acetyl-CoA carboxylase and hydroxymethylglutaryl-CoA reductase and mediates responses of these pathways to cellular stresses such as heat shock and depletion of glucose and ATP. AMPK is a heterotimeric complex comprised of a catalytic alpha subunit and two non-catalytic beta and gamma subunits that are believed to regulate the activity of the alpha subunit. Subunits of AMPK have a much wider distribution in non-lipogenic tissues such as brain, heart, spleen, and lung than expected. This distribution suggests that its role may extend beyond regulation of lipid metabolism alone.

The mitogen-activated protein kinases (MAP) are also members of the STK family. MAP kinases also regulate intracellular signaling pathways. They mediate signal transduction from the cell surface to the nucleus via phosphorylation cascades. Several subgroups have been identified, and each manifests different substrate specificities and responds to distinct extracellular stimuli (Egan, S. E. and Weinberg, R. A. (1993) *Nature* 365:781–783). MAP kinase signaling pathways are present in mammalian cells as well as in yeast. The extracellular stimuli that activate mammalian pathways include epidermal growth factor (EGF), ultraviolet light, hyperosmolar medium, heat shock, endotoxic lipopolysaccharide (LPS), and pro-inflammatory cytokines such as tumor necrosis factor (TNF) and interleukin-1 (IL-1).

PRK (proliferation-related kinase) is a serum/cytokine inducible STK that is involved in regulation of the cell cycle and cell proliferation in human megakaroytic cells (Li, B. et al. (1996) *J. Biol. Chem.* 271:19402–8). PRK is related to the polo (derived from humans polo gene) family of STKs implicated in cell division. PRK is downregulated in lung tumor tissue and may be a proto-oncogene whose deregulated expression in normal tissue leads to oncogenic transformation. Altered MAP kinase expression is implicated in a variety of disease conditions including cancer, inflammation, immune disorders, and disorders affecting growth and development.

The cyclin-dependent protein kinases (CDKs) are another group of STKs that control the progression of cells through the cell cycle. Cyclins are small regulatory proteins that act by binding to and activating CDKs that then trigger various phases of the cell cycle by phosphorylating and activating selected proteins involved in the mitotic process. CDKs are unique in that they require multiple inputs to become activated. In addition to the binding of cyclin, CDK activation requires the phosphorylation of a specific threonine residue and the dephosphorylation of a specific tyrosine residue.

Protein tyrosine kinases, PTKs, specifically phosphorylate tyrosine residues on their target proteins and may be divided into transmembrane, receptor PTKs and nontransmembrane, non-receptor PTKs. Transmembrane protein-tyrosine kinases are receptors for most growth factors. Binding of growth factor to the receptor activates the transfer of a phosphate group from ATP to selected tyrosine side chains of the receptor and other specific proteins. Growth factors (GF) associated with receptor PTKs include; epidermal GF, platelet-derived GF, fibroblast GF, hepatocyte GF, insulin and insulin-like GFs, nerve GF, vascular endothelial GF, and macrophage colony stimulating factor.

Non-receptor PTKs lack transmembrane regions and, instead, form complexes with the intracellular regions of cell surface receptors. Such receptors that function through non-receptor PTKs include those for cytokines, hormones (growth hormone and prolactin) and antigen-specific receptors on T and B lymphocytes.

Many of these PTKs were first identified as the products of mutant oncogenes in cancer cells where their activation was no longer subject to normal cellular controls. In fact, about one third of the known oncogenes encode PTKs, and it is well known that cellular transformation (oncogenesis) is often accompanied by increased tyrosine phosphorylation activity (Carbonneau H and Tonks NK (1992) *Annu. Rev. Cell. Biol.* 8:463–93). Regulation of PTK activity may therefore be an important strategy in controlling some types of cancer.

LIM Domain Kinases

The novel human protein, and encoding gene, provided by the present invention is related to the family of serine/threonine kinases in general, particularly LIM domain kinases (LIMK), and shows the highest degree of similarity to LIMK2, and the LIMK2b isoforn (Genbank gi8051618) in particular (see the amino acid sequence alignment of the protein of the present invention against LIMK2b provided in FIG. 2). LIMK proteins generally have serine/threonine kinase activity. The protein of the present invention may be a novel alternative splice form of the art-known protein provided in Genbank gi805161 ; however, the structure of the gene provided by the present invention is different from the art-known gene of gi8051618 and the first exon of the gene of the present invention is novel, suggesting a novel gene rather than an alternative splice form. Furthermore, the protein of the present invention lacks an LIM domain relative to gi8051618. The protein of the present invention does contain the kinase catalytic domain.

Approximately 40 LIM proteins, named for the LIM domains they contain, are known to exist in eukaryotes. LIM domains are conserved, cystein-rich structures that contain 2 zinc fingers that are thought to modulate protein-protein interactions. LIMK1 and LIMK2 are members of a LIM subfamily characterized by 2 N-terminal LIM domains and a C-terminal protein kinase domain. LIMK1 and LIMK2 mRNA expression varies greatly between different tissues. The protein kinase domains of LIMK1 and LIMK2 contain a unique sequence motif comprising Asp-Leu-Asn-Ser-His-Asn in subdomain VIB and a strongly basic insert between subdomains VII and VIII (Okano et al., *J. Biol. Chem.* 270 (52), 31321–31330 (1995)). The protein kinase domain present in LIMKs is significantly different than other kinase domains, sharing about 32% identity.

LIMK is activated by ROCK (a downstream effector of Rho) via phosphorylation. LIMK then phosphorylates cofilin, which inhibits its actin-depolymerizing activity, thereby leading to Rho-induced reorganization of the actin cytoskeleton (Maekawa et al., *Science* 285: 895–898, 1999).

The LIMK2a and LIMK2b alternative transcript forms are differentially expressed in a tissue-specific manner and are generated by variation in transcriptional initiation utilizing alternative promoters. LIMK2a contains 2 LIM domains, a PDZ domain (a domain that functions in protein-protein interactions targeting the protein to the submembranous compartment), and a kinase domain; whereas LIMK2b just has 1.5 LIM domains. Alteration of LIMK2a and LIMK2b regulation has been observed in some cancer cell lines (Osada et al., *Biochem. Biophys. Res. Commun.* 229: 582–589, 1996).

For a further review of LIMK proteins, see Nomoto et at, *Gene* 236 (2), 259–271 (1999).

Kinase proteins, particularly members of the serine/threonine kinase subfamily, are a major target for drug action and development. Accordingly, it is valuable to the field of pharmaceutical development to identify and characterize previously unknown members of this subfamily of kinase proteins. The present invention advances the state of the art by providing previously unidentified human kinase proteins that have homology to members of the serine/threonine kinase subfamily.

SUMMARY OF THE INVENTION

The present invention is based in part on the identification of amino acid sequences of human kinase peptides and proteins that are related to the serine/threonine kinase subfamily, as well as allelic variants and other mammalian orthologs thereof. These unique peptide sequences, and nucleic acid sequences that encode these peptides, can be used as models for the development of human therapeutic targets, aid in the identification of therapeutic proteins, and serve as targets for the development of human therapeutic agents that modulate kinase activity in cells and tissues that express the kinase. Experimental data as provided in FIG. 1

indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland.

DESCRIPTION OF THE FIGURE SHEETS

FIG. 1 provides the nucleotide sequence of a cDNA molecule that encodes the kinase protein of the present invention. (SEQ ID NO:1) In addition, structure and functional information is provided, such as ATG start, stop and tissue distribution, where available, that allows one to readily determine specific uses of inventions based on this molecular sequence. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland.

FIG. 2 provides the predicted amino acid sequence of the kinase of the present invention. (SEQ ID NO:2) In addition structure and functional information such as protein family, function, and modification sites is provided where available, allowing one to readily determine specific uses of inventions based on this molecular sequence.

FIG. 3 provides genomic sequences that span the gene encoding the kinase protein of the present invention. (SEQ ID NO:3) In addition structure and functional information, such as intron/exon structure, promoter location, etc., is provided where available, allowing one to readily determine specific uses of inventions based on this molecular sequence. As illustrated in FIG. 3, SNPs were identified at 42 different nucleotide positions.

DETAILED DESCRIPTION OF THE INVENTION

General Description

The present invention is based on the sequencing of the human genome. During the sequencing and assembly of the human genome, analysis of the sequence information revealed previously unidentified fragments of the human genome that encode peptides that share structural and/or sequence homology to protein/peptide/domains identified and characterized within the art as being a kinase protein or part of a kinase protein and are related to the serine/threonine kinase subfamily. Utilizing these sequences, additional genomic sequences were assembled and transcript and/or cDNA sequences were isolated and characterized. Based on this analysis, the present invention provides amino acid sequences of human kinase peptides and proteins that are related to the serine/threonine kinase subfamily, nucleic acid sequences in the form of transcript sequences, cDNA sequences and/or genomic sequences that encode these kinase peptides and proteins, nucleic acid variation (allelic information), tissue distribution of expression, and information about the closest art known protein/peptide/domain that has structural or sequence homology to the kinase of the present invention.

In addition to being previously unknown, the peptides that are provided in the present invention are selected based on their ability to be used for the development of commercially important products and services. Specifically, the present peptides are selected based on homology and/or structural relatedness to known kinase proteins of the serine/threonine kinase subfamily and the expression pattern observed. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. The art has clearly established the commercial importance of members of this family of proteins and proteins that have expression patterns similar to that of the present gene. Some of the more specific features of the peptides of the present invention, and the uses thereof, are described herein, particularly in the Background of the Invention and in the annotation provided in the Figures, and/or are known within the art for each of the known serine/threonine kinase family or subfamily of kinase proteins.

Specific Embodiments

Peptide Molecules

The present invention provides nucleic acid sequences that encode protein molecules that have been identified as being members of the kinase family of proteins and are related to the serine/threonine kinase subfamily (protein sequences are provided in FIG. 2, transcript/cDNA sequences are provided in FIG. 1 and genomic sequences are provided in FIG. 3). The peptide sequences provided in FIG. 2, as well as the obvious variants described herein, particularly allelic variants as identified herein and using the information in FIG. 3, will be referred herein as the kinase peptides of the present invention, kinase peptides, or peptides/proteins of the present invention.

The present invention provides isolated peptide and protein molecules that consist of, consist essentially of, or comprise the amino acid sequences of the kinase peptides disclosed in the FIG. 2, (encoded by the nucleic acid molecule shown in FIG. 1, transcript/cDNA or FIG. 3, genomic sequence), as well as all obvious variants of these peptides that are within the art to make and use. Some of these variants are described in detail below.

As used herein, a peptide is said to be "isolated" or "purified" when it is substantially free of cellular material or free of chemical precursors or other chemicals. The peptides of the present invention can be purified to homogeneity or other degrees of purity. The level of purification will be based on the intended use. The critical feature is that the preparation allows for the desired function of the peptide, even if in the presence of considerable amounts of other components (the features of an isolated nucleic acid molecule is discussed below).

In some uses, "substantially free of cellular material" includes preparations of the peptide having less than about 30% (by dry weight) other proteins (i.e., contaminating protein), less than about 20% other proteins, less than about 10% other proteins, or less than about 5% other proteins. When the peptide is recombinantly produced, it can also be substantially free of culture medium, i.e., culture medium represents less than about 20% of the volume of the protein preparation.

The language "substantially free of chemical precursors or other chemicals" includes preparations of the peptide in which it is separated from chemical precursors or other chemicals that are involved in its synthesis. In one embodiment, the language "substantially free of chemical precursors or other chemicals" includes preparations of the kinase peptide having less than about 30% (by dry weight) chemical precursors or other chemicals, less than about 20% chemical precursors or other chemicals, less than about 10% chemical precursors or other chemicals, or less than about 5% chemical precursors or other chemicals.

The isolated kinase peptide can be purified from cells that naturally express it, purified from cells that have been altered to express it (recombinant), or synthesized using known protein synthesis methods. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. For example, a nucleic acid molecule encoding the kinase peptide is cloned into an expression vector, the expression vector introduced into a host cell and the protein expressed in the host cell. The protein can then be isolated from the cells by an appropriate purification scheme using standard protein purification techniques. Many of these techniques are described in detail below.

Accordingly, the present invention provides proteins that consist of the amino acid sequences provided in FIG. 2 (SEQ ID NO:2), for example, proteins encoded by the transcript/cDNA nucleic acid sequences shown in FIG. 1 (SEQ ID NO:1) and the genornic sequences provided in FIG. 3 (SEQ ID NO:3). The amino acid sequence of such a protein is provided in FIG. 2. A protein consists of an amino acid sequence when the amino acid sequence is the final amino acid sequence of the protein.

The present invention further provides proteins that consist essentially of the amino acid sequences provided in FIG. 2 (SEQ ID NO:2), for example, proteins encoded by the transcript/cDNA nucleic acid sequences shown in FIG. 1 (SEQ ID NO:1) and the genomic sequences provided in FIG. 3 (SEQ ID NO:3). A protein consists essentially of an amino acid sequence when such an amino acid sequence is present with only a few additional amino acid residues, for example from about 1 to about 100 or so additional residues, typically from 1 to about 20 additional residues in the final protein.

The present invention further provides proteins that comprise the amino acid sequences provided in FIG. 2 (SEQ ID NO:2), for example, proteins encoded by the transcript/cDNA nucleic acid sequences shown in FIG. 1 (SEQ ID NO:1) and the genomic sequences provided in FIG. 3 (SEQ ID NO:3). A protein comprises an amino acid sequence when the amino acid sequence is at least part of the final amino acid sequence of the protein. In such a fashion, the protein can be only the peptide or have additional amino acid molecules, such as amino acid residues (contiguous encoded sequence) that are naturally associated with it or heterologous amino acid residues/peptide sequences. Such a protein can have a few additional amino acid residues or can comprise several hundred or more additional amino acids. The preferred classes of proteins that are comprised of the kinase peptides of the present invention are the naturally occurring mature proteins. A brief description of how various types of these proteins can be made/isolated is provided below.

The kinase peptides of the present invention can be attached to heterologous sequences to form chimeric or fusion proteins. Such chimeric and fusion proteins comprise a kinase peptide operatively linked to a heterologous protein having an amino acid sequence not substantially homologous to the kinase peptide. "Operatively linked" indicates that the kinase peptide and the heterologous protein are fused in-frame. The heterologous protein can be fused to the N-terminus or C-terminus of the kinase peptide.

In some uses, the fusion protein does not affect the activity of the kinase peptide per se. For example, the fusion protein can include, but is not limited to, enzymatic fusion proteins, for example beta-galactosidase fusions, yeast two-hybrid GAL fusions, poly-His fusions, MYC-tagged, HI-tagged and Ig fusions. Such fusion proteins, particularly poly-His fusions, can facilitate the purification of recombinant kinase peptide. In certain host cells (e.g., mammalian host cells), expression and/or secretion of a protein can be increased by using a heterologous signal sequence.

A chimeric or fusion protein can be produced by standard recombinant DNA techniques. For example, DNA fragments coding for the different protein sequences are ligated together in-frame in accordance with conventional techniques. In another embodiment, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers which give rise to complementary overhangs between two consecutive gene fragments which can subsequently be annealed and re-amplified to generate a chimeric gene sequence (see Ausubel et al., *Current Protocols in Molecular Biology*, 1992). Moreover, many expression vectors are commercially available that already encode a fusion moiety (e.g., a GST protein). A kinase peptide-encoding nucleic acid can be cloned into such an expression vector such that the fusion moiety is linked in-frame to the kinase peptide.

As mentioned above, the present invention also provides and enables obvious variants of the amino acid sequence of the proteins of the present invention, such as naturally occurring mature forms of the peptide, allelic/sequence variants of the peptides, non-naturally occurring recombinantly derived variants of the peptides, and orthologs and paralogs of the peptides. Such variants can readily be generated using art-known techniques in the fields of recombinant nucleic acid technology and protein biochemistry. It is understood, however, that variants exclude any amino acid sequences disclosed prior to the invention.

Such variants can readily be identified/made using molecular techniques and the sequence information disclosed herein. Further, such variants can readily be distinguished from other peptides based on sequence and/or structural homology to the kinase peptides of the present invention. The degree of homology/identity present will be based primarily on whether the peptide is a functional variant or non-functional variant, the amount of divergence present in the paralog family and the evolutionary distance between the orthologs.

To determine the percent identity of two amino acid sequences or two nucleic acid sequences, the sequences are aligned for optimal comparison purposes (e.g., gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be disregarded for comparison purposes). In a preferred embodiment, at least 30%, 40%, 50%, 60%, 70%, 80%, or 90% or more of the length of a reference sequence is aligned for comparison purposes. The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, then the molecules are identical at that position (as used herein amino acid or nucleic acid "identity" is equivalent to amino acid or nucleic acid "homology"). The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, which need to be introduced for optimal alignment of the two sequences.

The comparison of sequences and determination of percent identity and similarity between two sequences can be accomplished using a mathematical algorithm. (*Computational Molecular Biology*, Lesk, A. M., ed., Oxford University Press, New York, 1988; *Biocomputing: Informatics and Genome Projects*, Smith, D. W., ed., Academic Press, New York, 1993; *Computer Analysis of Sequence Data, Part* 1, Griffin, A. M., and Griffin, H. G., eds., Humana Press, New Jersey, 1994; *Sequence Analysis in Molecular Biology*, von Heinje, G., Academic Press, 1987; and *Sequence Analysis Primer*, Gribskov, M. and Devereux, J., eds., M Stockton Press, New York, 1991). In a preferred embodiment, the percent identity between two amino acid sequences is determined using the Needleman and Wunsch (*J. Mol. Biol.* (48):444–453 (1970)) algorithm which has been incorporated into the GAP program in the GCG software package (available at http://www.gcg.com), using either a Blossom 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6. In yet another preferred embodiment, the percent identity between two nucleotide sequences is determined using the GAP program in the GCG software package (Devereux, J., et al., *Nucleic Acids Res.* 12(1):387 (1984)) (available at http://www.gcg.com), using a NWS-gapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. In another embodiment, the percent identity between two amino acid or nucleotide sequences is determined using the algorithm of E. Myers and W. Miller (CABIOS, 4:11–17 (1989)) which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12 and a gap penalty of 4.

The nucleic acid and protein sequences of the present invention can further be used as a "query sequence" to perform a search against sequence databases to, for example, identify other family members or related sequences. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul, et al. (*J. Mol. Biol.* 215:403–10 (1990)). BLAST nucleotide searches can be performed with the NBLAST program, score=100, wordlength=12 to obtain nucleotide sequences homologous to the nucleic acid molecules of the invention. BLAST protein searches can be performed with the XBLAST program, score=50, wordlength=3 to obtain amino acid sequences homologous to the proteins of the invention. To obtain gapped alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul et al. (*Nucleic Acids Res.* 25(17):3389–3402 (1997)). When utilizing BLAST and gapped BLAST programs, the default parameters of the respective programs (e.g., XBLAST and NBLAST) can be used.

Full-length pre-processed forms, as well as mature processed forms, of proteins that comprise one of the peptides of the present invention can readily be identified as having complete sequence identity to one of the kinase peptides of the present invention as well as being encoded by the same genetic locus as the kinase peptide provided herein. The gene encoding the novel kinase protein of the present invention is located on a genome component that has been mapped to human chromosome 22 (as indicated in FIG. 3), which is supported by multiple lines of evidence, such as STS and BAC map data.

Allelic variants of a kinase peptide can readily be identified as being a human protein having a high degree (significant) of sequence homology/identity to at least a portion of the kinase peptide as well as being encoded by the same genetic locus as the kinase peptide provided herein. Genetic locus can readily be determined based on the genomic information provided in FIG. 3, such as the genomic sequence mapped to the reference human. The gene encoding the novel kinase protein of the present invention is located on a genome component that has been mapped to human chromosome 22 (as indicated in FIG. 3), which is supported by multiple lines of evidence, such as STS and BAC map data. As used herein, two proteins (or a region of the proteins) have significant homology when the amino acid sequences are typically at least about 70–80%, 80–90%, and more typically at least about 90–95% or more homologous. A significantly homologous amino acid sequence, according to the present invention, will be encoded by a nucleic acid sequence that will hybridize to a kinase peptide encoding nucleic acid molecule under stringent conditions as more fully described below.

FIG. 3 provides information on SNPs that have been found in the gene encoding the kinase protein of the present invention. SNPs were identified at 42 different nucleotide positions. Some of these SNPs, which are located outside the ORF and in introns, may affect gene transcription.

Paralogs of a kinase peptide can readily be identified as having some degree of significant sequence homology/identity to at least a portion of the kinase peptide, as being encoded by a gene from humans, and as having similar activity or function. Two proteins will typically be considered paralogs when the amino acid sequences are typically at least about 60% or greater, and more typically at least about 70% or greater homology through a given region or domain. Such paralogs will be encoded by a nucleic acid sequence that will hybridize to a kinase peptide encoding nucleic acid molecule under moderate to stringent conditions as more fully described below.

Orthologs of a kinase peptide can readily be identified as having some degree of significant sequence homology/identity to at least a portion of the kinase peptide as well as being encoded by a gene from another organism. Preferred orthologs will be isolated from mammals, preferably primates, for the development of human therapeutic targets and agents. Such orthologs will be encoded by a nucleic acid sequence that will hybridize to a kinase peptide encoding nucleic acid molecule under moderate to stringent conditions, as more fully described below, depending on the degree of relatedness of the two organisms yielding the proteins.

Non-naturally occurring variants of the kinase peptides of the present invention can readily be generated using recombinant techniques. Such variants include, but are not limited to deletions, additions and substitutions in the amino acid sequence of the kinase peptide. For example, one class of substitutions are conserved amino acid substitution. Such substitutions are those that substitute a given amino acid in a kinase peptide by another amino acid of like characteristics. Typically seen as conservative substitutions are the replacements, one for another, among the aliphatic amino acids Ala, Val, Leu, and Ile; interchange of the hydroxyl residues Ser and Thr; exchange of the acidic residues Asp and Glu; substitution between the amide residues Asn and Gln; exchange of the basic residues Lys and Arg; and replacements among the aromatic residues Phe and Tyr. Guidance concerning which amino acid changes are likely to be phenotypically silent are found in Bowie et al., *Science* 247:1306–1310 (1990).

Variant kinase peptides can be fully functional or can lack function in one or more activities, e.g. ability to bind substrate, ability to phosphorylate substrate, ability to mediate signaling, etc. Fully functional variants typically contain only conservative variation or variation in non-critical residues or in non-critical regions. FIG. 2 provides the result of protein analysis and can be used to identify critical domains/regions. Functional variants can also contain substitution of similar amino acids that result in no change or an insignificant change in function. Alternatively, such substitutions may positively or negatively affect function to some degree.

Non-functional variants typically contain one or more non-conservative amino acid substitutions, deletions, insertions, inversions, or truncation or a substitution, insertion, inversion, or deletion in a critical residue or critical region.

Amino acids that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (Cunningham et al., *Science* 244:1081–1085 (1989)), particularly using the results provided in FIG. 2. The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity such as kinase activity or in assays such as an in vitro proliferative activity. Sites that are critical for binding partner/substrate binding can also be determined by structural analysis such as crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith et al., *J. Mol. Biol.* 224:899–904 (1992); de Vos et al. *Science* 255:306–312 (1992)).

The present invention further provides fragments of the kinase peptides, in addition to proteins and peptides that comprise and consist of such fragments, particularly those comprising the residues identified in FIG. 2. The fragments to which the invention pertains, however, are not to be construed as encompassing fragments that may be disclosed publicly prior to the present invention.

As used herein, a fragment comprises at least 8, 10, 12, 14, 16, or more contiguous amino acid residues from a kinase peptide. Such fragments can be chosen based on the ability to retain one or more of the biological activities of the kinase peptide or could be chosen for the ability to perform a function, e.g. bind a substrate or act as an immunogen. Particularly important fragments are biologically active fragments, peptides that are, for example, about 8 or more amino acids in length. Such fragments will typically comprise a domain or motif of the kinase peptide, e.g., active site, a transmembrane domain or a substrate-binding domain. Further, possible fragments include, but are not limited to, domain or motif containing fragments, soluble peptide fragments, and fragments containing immunogenic structures. Predicted domains and functional sites are readily identifiable by computer programs well known and readily available to those of skill in the art (e.g., PROSITE analysis). The results of one such analysis are provided in FIG. 2.

Polypeptides often contain amino acids other than the 20 amino acids commonly referred to as the 20 naturally occurring amino acids. Further, many amino acids, including the terminal amino acids, may be modified by natural processes, such as processing and other post-translational modifications, or by chemical modification techniques well known in the art. Common modifications that occur naturally in kinase peptides are described in basic texts, detailed monographs, and the research literature, and they are well known to those of skill in the art (some of these features are identified in FIG. 2).

Known modifications include, but are not limited to, acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphotidylinositol, cross-linking, cyclization, disulfide bond formation, demethylation, formation of covalent crosslinks, formation of cystine, formation of pyroglutamate, formylation, gamma carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination.

Such modifications are well known to those of skill in the art and have been described in great detail in the scientific literature. Several particularly common modifications, glycosylation, lipid attachment, sulfation, gamma-carboxylation of glutamic acid residues, hydroxylation and ADP-ribosylation, for instance, are described in most basic texts, such as *Proteins—Structure and Molecular Properties*, 2nd Ed., T. E. Creighton, W. H. Freeman and Company, New York (1993). Many detailed reviews are available on this subject, such as by Wold, F., *Posttranslational Covalent Modification of Proteins*, B. C. Johnson, Ed., Academic Press, New York 1–12 (1983); Seifter et al. (*Meth. Enzymol.* 182: 626–646 (1990)) and Rattan et al. (*Ann. N.Y. Acad. Sci.* 663:48–62 (1992)).

Accordingly, the kinase peptides of the present invention also encompass derivatives or analogs in which a substituted amino acid residue is not one encoded by the genetic code, in which a substituent group is included, in which the mature kinase peptide is fused with another compound, such as a compound to increase the half-life of the kinase peptide (for example, polyethylene glycol), or in which the additional amino acids are fused to the mature kinase peptide, such as a leader or secretory sequence or a sequence for purification of the mature kinase peptide or a pro-protein sequence.

Protein/Peptide Uses

The proteins of the present invention can be used in substantial and specific assays related to the functional information provided in the Figures; to raise antibodies or to elicit another immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the protein (or its binding partner or ligand) in biological fluids; and as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state). Where the protein binds or potentially binds to another protein or ligand (such as, for example, in a kinase-effector protein interaction or kinase-ligand interaction), the protein can be used to identify the binding partner/ligand so as to develop a system to identify inhibitors of the binding interaction. Any or all of these uses are capable of being developed into reagent grade or kit format for commercialization as commercial products.

Methods for performing the uses listed above are well known to those skilled in the art. References disclosing such methods include "Molecular Cloning: A Laboratory Manual", 2d ed., Cold Spring Harbor Laboratory Press, Sambrook, J., E. F. Fritsch and T. Maniatis eds., 1989, and "Methods in Enzymology: Guide to Molecular Cloning Techniques", Academic Press, Berger, S. L. and A. R. Kimmel eds., 1987.

The potential uses of the peptides of the present invention are based primarily on the source of the protein as well as the class/action of the protein. For example, kinases isolated from humans and their human/mammalian orthologs serve as targets for identifying agents for use in mammalian therapeutic applications, e.g. a human drug, particularly in modulating a biological or pathological response in a cell or tissue that expresses the kinase. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain. A large percentage of pharmaceutical agents are being developed that modulate the activity of kinase proteins, particularly members of the serine/threonine kinase subfamily (see Background of the Invention). The structural and functional information provided in the Background and Figures provide specific and substantial uses for the molecules of the present invention, particularly in combination with the expression information provided in FIG. 1. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. Such uses can readily be determined using the information provided herein, that which is known in the art, and routine experimentation.

The proteins of the present invention (including variants and fragments that may have been disclosed prior to the present invention) are useful for biological assays related to kinases that are related to members of the serine/threonine kinase subfamily. Such assays involve any of the known kinase functions or activities or properties useful for diagnosis and treatment of kinase-related conditions that are specific for the subfamily of kinases that the one of the present invention belongs to, particularly in cells and tissues that express the kinase. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain.

The proteins of the present invention are also usefull in drug screening assays, in cell-based or cell-free systems. Cell-based systems can be native, i.e., cells that normally express the kinase, as a biopsy or expanded in cell culture. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. In an alternate embodiment, cell-based assays involve recombinant host cells expressing the kinase protein.

The polypeptides can be used to identify compounds that modulate kinase activity of the protein in its natural state or an altered form that causes a specific disease or pathology associated with the kinase. Both the kinases of the present invention and appropriate variants and fragments can be used in high-throughput screens to assay candidate compounds for the ability to bind to the kinase. These compounds can be further screened against a functional kinase to determine the effect of the compound on the kinase activity. Further, these compounds can be tested in animal or invertebrate systems to determine activity/effectiveness. Compounds can be identified that activate (agonist) or inactivate (antagonist) the kinase to a desired degree.

Further, the proteins of the present invention can be used to screen a compound for the ability to stimulate or inhibit interaction between the kinase protein and a molecule that normally interacts with the kinase protein, e.g. a substrate or a component of the signal pathway that the kinase protein normally interacts (for example, another kinase). Such assays typically include the steps of combining the kinase protein with a candidate compound under conditions that allow the kinase protein, or fragment, to interact with the target molecule, and to detect the formation of a complex between the protein and the target or to detect the biochemical consequence of the interaction with the kinase protein and the target, such as any of the associated effects of signal transduction such as protein phosphorylation, cAMP turnover, and adenylate cyclase activation, etc.

Candidate compounds include, for example, 1) peptides such as soluble peptides, including Ig-tailed fusion peptides and members of random peptide libraries (see, e.g., Lam et al., *Nature* 354:82–84 (1991); Houghten et al., *Nature* 354:84–86 (1991)) and combinatorial chemistry-derived molecular libraries made of D- and/or L-configuration amino acids; 2) phosphopeptides (e.g., members of random and partially degenerate, directed phosphopeptide libraries, see, e.g., Songyang et al., *Cell* 72:767–778 (1993)); 3) antibodies (e.g., polyclonal, monoclonal, humanized, anti-idiotypic, chimeric, and single chain antibodies as well as Fab, $F(ab')_2$, Fab expression library fragments, and epitope-binding fragments of antibodies); and 4) small organic and inorganic molecules (e.g., molecules obtained from combinatorial and natural product libraries).

One candidate compound is a soluble fragment of the receptor that competes for substrate binding. Other candidate compounds include mutant kinases or appropriate fragments containing mutations that affect kinase function and thus compete for substrate. Accordingly, a fragment that competes for substrate, for example with a higher affinity, or a fragment that binds substrate but does not allow release, is encompassed by the invention.

The invention further includes other end point assays to identify compounds that modulate (stimulate or inhibit) kinase activity. The assays typically involve an assay of events in the signal transduction pathway that indicate kinase activity. Thus, the phosphorylation of a substrate, activation of a protein, a change in the expression of genes that are up- or down-regulated in response to the kinase protein dependent signal cascade can be assayed.

Any of the biological or biochemical functions mediated by the kinase can be used as an endpoint assay. These include all of the biochemical or biochemical/biological events described herein, in the references cited herein, incorporated by reference for these endpoint assay targets, and other functions known to those of ordinary skill in the art or that can be readily identified using the information provided in the Figures, particularly FIG. 2. Specifically, a biological function of a cell or tissues that expresses the kinase can be assayed. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain.

Binding and/or activating compounds can also be screened by using chimeric kinase proteins in which the amino terminal extracellular domain, or parts thereof, the entire transmembrane domain or subregions, such as any of the seven transmembrane segments or any of the intracellular or extracellular loops and the carboxy terminal intracellular domain, or parts thereof, can be replaced by heterologous domains or subregions. For example, a substrate-binding region can be used that interacts with a different substrate then that which is recognized by the native kinase. Accordingly, a different set of signal transduction components is available as an end-point assay for activation. This allows for assays to be performed in other than the specific host cell from which the kinase is derived.

The proteins of the present invention are also useful in competition binding assays in methods designed to discover compounds that interact with the kinase (e.g. binding partners and/or ligands). Thus, a compound is exposed to a kinase polypeptide under conditions that allow the compound to bind or to otherwise interact with the polypeptide. Soluble kinase polypeptide is also added to the mixture. If the test compound interacts with the soluble kinase polypeptide, it decreases the amount of complex formed or activity from the kinase target. This type of assay is particularly useful in cases in which compounds are sought that interact with specific regions of the kinase. Thus, the soluble polypeptide that competes with the target kinase region is designed to contain peptide sequences corresponding to the region of interest.

To perform cell free drug screening assays, it is sometimes desirable to immobilize either the kinase protein, or fragment, or its target molecule to facilitate separation of complexes from uncomplexed forms of one or both of the proteins, as well as to accommodate automation of the assay.

Techniques for immobilizing proteins on matrices can be used in the drug screening assays. In one embodiment, a fusion protein can be provided which adds a domain that allows the protein to be bound to a matrix. For example, glutathione-S-transferase fusion proteins can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, Mo.) or glutathione derivatized microtitre plates, which are then combined with the cell lysates (e.g., ^{35}S-labeled) and the candidate compound, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads are washed to remove any unbound label, and the matrix immobilized and radiolabel determined directly, or in the supernatant after the complexes are dissociated. Alternatively, the complexes can be dissociated from the matrix, separated by SDS-PAGE, and the level of kinase-binding protein found in the bead fraction quantitated from the gel using standard electrophoretic techniques. For example, either the polypeptide or its target molecule can be immobilized utilizing conjugation of biotin and streptavidin using techniques well known in the art. Alternatively, antibodies reactive with the protein but which do not interfere with binding of the protein to its target molecule can be derivatized to the wells of the plate, and the protein trapped in the wells by antibody conjugation. Preparations of a kinase-binding protein and a candidate compound are incubated in the kinase protein-presenting wells and the amount of complex trapped in the well can be quantitated. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the kinase protein target molecule, or which are reactive with kinase protein and compete with the target molecule, as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the target molecule.

Agents that modulate one of the kinases of the present invention can be identified using one or more of the above assays, alone or in combination. It is generally preferable to use a cell-based or cell free system first and then confirm activity in an animal or other model system. Such model systems are well known in the art and can readily be employed in this context.

Modulators of kinase protein activity identified according to these drug screening assays can be used to treat a subject with a disorder mediated by the kinase pathway, by treating cells or tissues that express the kinase. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. These methods of treatment include the steps of administering a modulator of kinase activity in a pharmaceutical composition to a subject in need of such treatment, the modulator being identified as described herein.

In yet another aspect of the invention, the kinase proteins can be used as "bait proteins" in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Pat. No. 5,283,317; Zervos et al. (1993) *Cell* 72:223–232; Madura et al. (1993) *J. Biol. Chem.* 268:12046–12054; Bartel et al. (1993) *Biotechniques* 14:920–924; Iwabuchi et al. (1993) *Oncogene* 8:1693/1696; and Brent WO94110300), to identify other proteins, which bind to or interact with the kinase and are involved in kinase activity. Such kinase-binding proteins are also likely to be involved in the propagation of signals by the kinase proteins or kinase targets as, for example, downstream elements of a kinase-mediated signaling pathway. Alternatively, such kinase-binding proteins are likely to be kinase inhibitors.

The two-hybrid system is based on the modular nature of most transcription factors, which consist of separable DNA-binding and activation domains. Briefly, the assay utilizes two different DNA constructs. In one construct, the gene that codes for a kinase protein is fused to a gene encoding the DNA binding domain of a known transcription factor (e.g., GAL-4). In the other construct, a DNA sequence, from a library of DNA sequences, that encodes an unidentified protein ("prey" or "sample") is fused to a gene that codes for the activation domain of the known transcription factor. If the "bait" and the "prey" proteins are able to interact, in vivo, forming a kinase-dependent complex, the DNA-binding and activation domains of the transcription factor are brought into close proximity. This proximity allows transcription of a reporter gene (e.g., LacZ) which is operably linked to a transcriptional regulatory site responsive to the transcription factor. Expression of the reporter gene can be detected and cell colonies containing the functional transcription factor can be isolated and used to obtain the cloned gene which encodes the protein which interacts with the kinase protein.

This invention further pertains to novel agents identified by the above-described screening assays. Accordingly, it is within the scope of this invention to further use an agent identified as described herein in an appropriate animal model. For example, an agent identified as described herein (e.g., a kinase-modulating agent, an antisense kinase nucleic acid molecule, a kinase-specific antibody, or a kinase-binding partner) can be used in an animal or other model to determine the efficacy, toxicity, or side effects of treatment with such an agent. Alternatively, an agent identified as described herein can be used in an animal or other model to determine the mechanism of action of such an agent. Furthermore, this invention pertains to uses of novel agents identified by the above-described screening assays for treatments as described herein.

The kinase proteins of the present invention are also useful to provide a target for diagnosing a disease or predisposition to disease mediated by the peptide. Accordingly, the invention provides methods for detecting the presence, or levels of, the protein (or encoding mRNA) in a cell, tissue, or organism. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. The method involves contacting a biological sample with a compound capable of interacting with the kinase protein such that the interaction can be detected. Such an assay can be provided in a single detection format or a multi-detection format such as an antibody chip array.

One agent for detecting a protein in a sample is an antibody capable of selectively binding to protein. A biological sample includes tissues, cells and biological fluids isolated from a subject, as well as tissues, cells and fluids present within a subject.

The peptides of the present invention also provide targets for diagnosing active protein activity, disease, or predisposition to disease, in a patient having a variant peptide, particularly activities and conditions that are known for other members of the family of proteins to which the present one belongs. Thus, the peptide can be isolated from a biological sample and assayed for the presence of a genetic mutation that results in aberrant peptide. This includes amino acid substitution, deletion, insertion, rearrangement, (as the result of aberrant splicing events), and inappropriate post-translational modification. Analytic methods include altered electrophoretic mobility, altered tryptic peptide digest, altered kinase activity in cell-based or cell-free assay, alteration in substrate or antibody-binding pattern, altered isoelectric point, direct amino acid sequencing, and any other of the known assay techniques useful for detecting mutations in a protein. Such an assay can be provided in a single detection format or a multi-detection format such as an antibody chip array.

In vitro techniques for detection of peptide include enzyme linked immunosorbent assays (ELISAs), Western blots, immunoprecipitations and immunofluorescence using a detection reagent, such as an antibody or protein binding agent. Alternatively, the peptide can be detected in vivo in a subject by introducing into the subject a labeled anti-peptide antibody or other types of detection agent. For example, the antibody can be labeled with a radioactive marker whose presence and location in a subject can be detected by standard imaging techniques. Particularly useful are methods that detect the allelic variant of a peptide expressed in a subject and methods which detect fragments of a peptide in a sample.

The peptides are also useful in pharmacogenomic analysis. Pharmacogenomics deal with clinically significant hereditary variations in the response to drugs due to altered drug disposition and abnormal action in affected persons. See, e.g., Eichelbaum, M. (*Clin. Exp. Pharmacol. Physiol.* 23(10–11):983–985 (1996)), and Linder, M. W. (*Clin. Chem.* 43(2):254–266 (1997)). The clinical outcomes of these variations result in severe toxicity of therapeutic drugs in certain individuals or therapeutic failure of drugs in certain individuals as a result of individual variation in metabolism. Thus, the genotype of the individual can determine the way a therapeutic compound acts on the body or the way the body metabolizes the compound. Further, the activity of drug metabolizing enzymes effects both the intensity and duration of drug action. Thus, the pharmacogenomics of the individual permit the selection of effective compounds and effective dosages of such compounds for prophylactic or therapeutic treatment based on the individual's genotype. The discovery of genetic polymorphisms in some drug metabolizing enzymes has explained why some patients do not obtain the expected drug effects, show an exaggerated drug effect, or experience serious toxicity from standard drug dosages. Polymorphisms can be expressed in the phenotype of the extensive metabolizer and the phenotype of the poor metabolizer. Accordingly, genetic polymorphism may lead to allelic protein variants of the kinase protein in which one or more of the kinase functions in one population is different from those in another population. The peptides thus allow a target to ascertain a genetic predisposition that can affect treatment modality. Thus, in a ligand-based treatment, polymorphism may give rise to amino terminal extracellular domains and/or other substrate-binding regions that are more or less active in substrate binding, and kinase activation. Accordingly, substrate dosage would necessarily be modified to maximize the therapeutic effect within a given population containing a polymorphism. As an alternative to genotyping, specific polymorphic peptides could be identified.

The peptides are also useful for treating a disorder characterized by an absence of, inappropriate, or unwanted expression of the protein. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. Accordingly, methods for treatment include the use of the kinase protein or fragments.

Antibodies

The invention also provides antibodies that selectively bind to one of the peptides of the present invention, a protein comprising such a peptide, as well as variants and fragments thereof. As used herein, an antibody selectively binds a target peptide when it binds the target peptide and does not significantly bind to unrelated proteins. An antibody is still considered to selectively bind a peptide even if it also binds to other proteins that are not substantially homologous with the target peptide so long as such proteins share homology with a fragment or domain of the peptide target of the antibody. In this case, it would be understood that antibody binding to the peptide is still selective despite some degree of cross-reactivity.

As used herein, an antibody is defined in terms consistent with that recognized within the art: they are multi-subunit proteins produced by a mammalian organism in response to an antigen challenge. The antibodies of the present invention include polyclonal antibodies and monoclonal antibodies, as well as fragments of such antibodies, including, but not limited to, Fab or F(ab')$_2$, and Fv fragments.

Many methods are known for generating and/or identifying antibodies to a given target peptide. Several such methods are described by Harlow, Antibodies, Cold Spring Harbor Press, (1989).

In general, to generate antibodies, an isolated peptide is used as an immunogen and is administered to a mammalian organism, such as a rat, rabbit or mouse. The full-length protein, an antigenic peptide fragment or a fusion protein can be used. Particularly important fragments are those covering functional domains, such as the domains identified in FIG. 2, and domain of sequence homology or divergence amongst the family, such as those that can readily be identified using protein alignment methods and as presented in the Figures.

Antibodies are preferably prepared from regions or discrete fragments of the kinase proteins. Antibodies can be prepared from any region of the peptide as described herein. However, preferred regions will include those involved in function/activity and/or kinase/binding partner interaction. FIG. 2 can be used to identify particularly important regions while sequence alignment can be used to identify conserved and unique sequence fragments.

An antigenic fragment will typically comprise at least 8 contiguous amino acid residues. The antigenic peptide can comprise, however, at least 10, 12, 14, 16 or more amino acid residues. Such fragments can be selected on a physical property, such as fragments correspond to regions that are located on the surface of the protein, e.g., hydrophilic regions or can be selected based on sequence uniqueness (see FIG. 2).

Detection on an antibody of the present invention can be facilitated by coupling (i.e., physically linking) the antibody to a detectable substance. Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, β-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I, ^{131}I, ^{35}S or ^{3}H.

Antibody Uses

The antibodies can be used to isolate one of the proteins of the present invention by standard techniques, such as affinity chromatography or immunoprecipitation. The antibodies can facilitate the purification of the natural protein from cells and recombinantly produced protein expressed in host cells. In addition, such antibodies are useful to detect the presence of one of the proteins of the present invention in cells or tissues to determine the pattern of expression of the protein among various tissues in an organism and over the course of normal development. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain. Further, such antibodies can be used to detect protein in situ, in vitro, or in a cell lysate or supernatant in order to evaluate the abundance and pattern of expression. Also, such antibodies can be used to assess abnormal tissue distribution or abnormal expression during development or progression of a biological condition. Antibody detection of circulating fragments of the full length protein can be used to identify turnover.

Further, the antibodies can be used to assess expression in disease states such as in active stages of the disease or in an individual with a predisposition toward disease related to the protein's function. When a disorder is caused by an inappropriate tissue distribution, developmental expression, level of expression of the protein, or expressed/processed form, the antibody can be prepared against the normal protein. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. If a disorder is characterized by a specific mutation in the protein, antibodies specific for this mutant protein can be used to assay for the presence of the specific mutant protein.

The antibodies can also be used to assess normal and aberrant subcellular localization of cells in the various tissues in an organism. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. The diagnostic uses can be applied, not only in genetic testing, but also in monitoring a treatment modality. Accordingly, where treatment is ultimately aimed at correcting expression level or the presence of aberrant sequence and aberrant tissue distribution or developmental expression, antibodies directed against the protein or relevant fragments can be used to monitor therapeutic efficacy.

Additionally, antibodies are useful in pharmacogenomic analysis. Thus, antibodies prepared against polymorphic proteins can be used to identify individuals that require modified treatment modalities. The antibodies are also useful as diagnostic tools as an immunological marker for aberrant protein analyzed by electrophoretic mobility, isoelectric point, tryptic peptide digest, and other physical assays known to those in the art.

The antibodies are also useful for tissue typing. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. Thus, where a specific protein has been correlated with expression in a specific tissue, antibodies that are specific for this protein can be used to identify a tissue type.

The antibodies are also useful for inhibiting protein function, for example, blocking the binding of the kinase peptide to a binding partner such as a substrate. These uses can also be applied in a therapeutic context in which treatment involves inhibiting the protein's function. An antibody can be used, for example, to block binding, thus modulating (agonizing or antagonizing) the peptides activity. Antibodies can be prepared against specific fragments containing sites required for function or against intact protein that is associated with a cell or cell membrane. See FIG. 2 for structural information relating to the proteins of the present invention.

The invention also encompasses kits for using antibodies to detect the presence of a protein in a biological sample. The kit can comprise antibodies such as a labeled or labelable antibody and a compound or agent for detecting protein in a biological sample; means for determining the amount of protein in the sample; means for comparing the amount of protein in the sample with a standard; and instructions for use. Such a kit can be supplied to detect a single protein or epitope or can be configured to detect one of a multitude of epitopes, such as in an antibody detection array. Arrays are described in detail below for nuleic acid arrays and similar methods have been developed for antibody arrays.

Nucleic Acid Molecules

The present invention further provides isolated nucleic acid molecules that encode a kinase peptide or protein of the present invention (cDNA, transcript and genomic sequence). Such nucleic acid molecules will consist of, consist essentially of, or comprise a nucleotide sequence that encodes one of the kinase peptides of the present invention, an allelic variant thereof, or an ortholog or paralog thereof.

As used herein, an "isolated" nucleic acid molecule is one that is separated from other nucleic acid present in the natural source of the nucleic acid. Preferably, an "isolated" nucleic acid is free of sequences which naturally flank the nucleic acid (i.e., sequences located at the 5' and 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. However, there can be some flanking nucleotide sequences, for example up to about 5KB, 4KB, 3KB, 2KB, or 1KB or less, particularly contiguous peptide encoding sequences and peptide encoding sequences within the same gene but separated by introns in the genomic sequence. The important point is that the nucleic acid is isolated from remote and unimportant flanking sequences such that it can be subjected to the specific manipulations described herein such as recombinant expression, preparation of probes and primers, and other uses specific to the nucleic acid sequences.

Moreover, an "isolated" nucleic acid molecule, such as a transcript/cDNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or chemical precursors or other chemicals when chemically synthesized. However, the nucleic acid molecule can be fused to other coding or regulatory sequences and still be considered isolated.

For example, recombinant DNA molecules contained in a vector are considered isolated. Further examples of isolated DNA molecules include recombinant DNA molecules maintained in heterologous host cells or purified (partially or substantially) DNA molecules in solution. Isolated RNA molecules include in vivo or in vitro RNA transcripts of the isolated DNA molecules of the present invention. Isolated nucleic acid molecules according to the present invention further include such molecules produced synthetically.

Accordingly, the present invention provides nucleic acid molecules that consist of the nucleotide sequence shown in FIG. 1 or 3 (SEQ ID NO:1, transcript sequence and SEQ ID NO:3, genomic sequence), or any nucleic acid molecule that encodes the protein provided in FIG. 2, SEQ ID NO:2. A nucleic acid molecule consists of a nucleotide sequence when the nucleotide sequence is the complete nucleotide sequence of the nucleic acid molecule.

The present invention further provides nucleic acid molecules that consist essentially of the nucleotide sequence shown in FIG. 1 or 3 (SEQ ID NO:1, transcript sequence and SEQ ID NO:3, genomic sequence), or any nucleic acid molecule that encodes the protein provided in FIG. 2, SEQ ID NO:2. A nucleic acid molecule consists essentially of a nucleotide sequence when such a nucleotide sequence is present with only a few additional nucleic acid residues in the final nucleic acid molecule.

The present invention further provides nucleic acid molecules that comprise the nucleotide sequences shown in FIG. 1 or 3 (SEQ ID NO:1, transcript sequence and SEQ ID NO:3, genomic sequence), or any nucleic acid molecule that encodes the protein provided in FIG. 2, SEQ ID NO:2. A nucleic acid molecule comprises a nucleotide sequence when the nucleotide sequence is at least part of the final nucleotide sequence of the nucleic acid molecule. In such a fashion, the nucleic acid molecule can be only the nucleotide sequence or have additional nucleic acid residues, such as nucleic acid residues that are naturally associated with it or heterologous nucleotide sequences. Such a nucleic acid molecule can have a few additional nucleotides or can comprises several hundred or more additional nucleotides. A brief description of how various types of these nucleic acid molecules can be readily made/isolated is provided below.

In FIGS. 1 and 3, both coding and non-coding sequences are provided. Because of the source of the present invention, humans genomic sequence (FIG. 3) and cDNA/transcript sequences (FIG. 1), the nucleic acid molecules in the Figures will contain genomic intronic sequences, 5' and 3' non-coding sequences, gene regulatory regions and non-coding intergenic sequences. In general such sequence features are either noted in FIGS. 1 and 3 or can readily be identified using computational tools known in the art. As discussed below, some of the non-coding regions, particularly gene regulatory elements such as promoters, are useful for a variety of purposes, e.g. control of heterologous gene expression, target for identifying gene activity modulating compounds, and are particularly claimed as fragments of the genomic sequence provided herein.

The isolated nucleic acid molecules can encode the mature protein plus additional amino or carboxyl-terminal amino acids, or amino acids interior to the mature peptide (when the mature form has more than one peptide chain, for instance). Such sequences may play a role in processing of a protein from precursor to a mature form, facilitate protein trafficking, prolong or shorten protein half-life or facilitate manipulation of a protein for assay or production, among other things. As generally is the case in situ, the additional amino acids may be processed away from the mature protein by cellular enzymes.

As mentioned above, the isolated nucleic acid molecules include, but are not limited to, the sequence encoding the kinase peptide alone, the sequence encoding the mature peptide and additional coding sequences, such as a leader or secretory sequence (e.g., a pre-pro or pro-protein sequence), the sequence encoding the mature peptide, with or without the additional coding sequences, plus additional non-coding sequences, for example introns and non-coding 5' and 3' sequences such as transcribed but non-translated sequences that play a role in transcription, mRNA processing (including splicing and polyadenylation signals), ribosome binding and stability of mRNA. In addition, the nucleic acid molecule may be fused to a marker sequence encoding, for example, a peptide that facilitates purification.

Isolated nucleic acid molecules can be in the form of RNA, such as mRNA, or in the form DNA, including cDNA and genomic DNA obtained by cloning or produced by chemical synthetic techniques or by a combination thereof. The nucleic acid, especially DNA, can be double-stranded or single-stranded. Single-stranded nucleic acid can be the coding strand (sense strand) or the non-coding strand (anti-sense strand).

The invention further provides nucleic acid molecules that encode fragments of the peptides of the present invention as well as nucleic acid molecules that encode obvious variants of the kinase proteins of the present invention that are described above. Such nucleic acid molecules may be naturally occurring, such as allelic variants (same locus), paralogs (different locus), and orthologs (different organism), or may be constructed by recombinant DNA methods or by chemical synthesis. Such non-naturally occurring variants may be made by mutagenesis techniques, including those applied to nucleic acid molecules, cells, or organisms. Accordingly, as discussed above, the variants can contain nucleotide substitutions, deletions, inversions and insertions. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions.

The present invention further provides non-coding fragments of the nucleic acid molecules provided in FIGS. 1 and 3. Preferred non-coding fragments include, but are not limited to, promoter sequences, enhancer sequences, gene modulating sequences and gene termination sequences. Such fragments are useful in controlling heterologous gene expression and in developing screens to identify gene-modulating agents. A promoter can readily be identified as being 5' to the ATG start site in the genomic sequence provided in FIG. 3.

A fragment comprises a contiguous nucleotide sequence greater than 12 or more nucleotides. Further, a fragment could at least 30, 40, 50, 100, 250 or 500 nucleotides in length. The length of the fragment will be based on its intended use. For example, the fragment can encode epitope bearing regions of the peptide, or can be useful as DNA probes and primers. Such fragments can be isolated using the known nucleotide sequence to synthesize an oligonucleotide probe. A labeled probe can then be used to screen a cDNA library, genomic DNA library, or mRNA to isolate nucleic acid corresponding to the coding region. Further, primers can be used in PCR reactions to clone specific regions of gene.

A probe/primer typically comprises substantially a purified oligonucleotide or oligonucleotide pair. The oligonucleotide typically comprises a region of nucleotide sequence that hybridizes under stringent conditions to at least about 12, 20, 25, 40, 50 or more consecutive nucleotides.

Orthologs, homologs, and allelic variants can be identified using methods well known in the art. As described in the Peptide Section, these variants comprise a nucleotide sequence encoding a peptide that is typically 60–70%, 70–80%, 80–90%, and more typically at least about 90–95% or more homologous to the nucleotide sequence shown in the Figure sheets or a fragment of this sequence. Such nucleic acid molecules can readily be identified as being able to hybridize under moderate to stringent conditions, to the nucleotide sequence shown in the Figure sheets or a fragment of the sequence. Allelic variants can readily be determined by genetic locus of the encoding gene. The gene encoding the novel kinase protein of the present invention is located on a genome component that has been mapped to human chromosome 22 (as indicated in FIG. 3), which is supported by multiple lines of evidence, such as STS and BAC map data.

FIG. 3 provides information on SNPs that have been found in the gene encoding the kinase protein of the present invention. SNPs were identified at 42 different nucleotide positions. Some of these SNPs, which are located outside the ORF and in introns, may affect gene transcription.

As used herein, the term "hybridizes under stringent conditions" is intended to describe conditions for hybridization and washing under which nucleotide sequences encoding a peptide at least 60–70% homologous to each other typically remain hybridized to each other. The conditions can be such that sequences at least about 60%, at least about 70%, or at least about 80% or more homologous to each other typically remain hybridized to each other. Such stringent conditions are known to those skilled in the art and can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. (1989), 6.3.1–6.3.6. One example of stringent hybridization conditions are hybridization in 6× sodium chloride/sodium citrate (SSC) at about 45C, followed by one or more washes in 0.2×SSC, 0.1% SDS at 50–65C. Examples of moderate to low stringency hybridization conditions are well known in the art.

Nucleic Acid Molecule Uses

The nucleic acid molecules of the present invention are useful for probes, primers, chemical intermediates, and in biological assays. The nucleic acid molecules are useful as a hybridization probe for messenger RNA, transcript/cDNA and genomic DNA to isolate full-length cDNA and genomic clones encoding the peptide described in FIG. 2 and to isolate cDNA and genomic clones that correspond to variants (alleles, orthologs, etc.) producing the same or related peptides shown in FIG. 2. As illustrated in FIG. 3, SNPs were identified at 42 different nucleotide positions.

The probe can correspond to any sequence along the entire length of the nucleic acid molecules provided in the Figures. Accordingly, it could be derived from 5' noncoding regions, the coding region, and 3' noncoding regions. However, as discussed, fragments are not to be construed as encompassing fragments disclosed prior to the present invention.

The nucleic acid molecules are also useful as primers for PCR to amplify any given region of a nucleic acid molecule and are useful to synthesize antisense molecules of desired length and sequence.

The nucleic acid molecules are also useful for constructing recombinant vectors. Such vectors include expression vectors that express a portion of, or all of, the peptide sequences. Vectors also include insertion vectors, used to integrate into another nucleic acid molecule sequence, such as into the cellular genome, to alter in situ expression of a gene and/or gene product. For example, an endogenous coding sequence can be replaced via homologous recombination with all or part of the coding region containing one or more specifically introduced mutations.

The nucleic acid molecules are also useful for expressing antigenic portions of the proteins.

The nucleic acid molecules are also useful as probes for determining the chromosomal positions of the nucleic acid molecules by means of in situ hybridization methods. The gene encoding the novel kinase protein of the present invention is located on a genome component that has been mapped to human chromosome 22 (as indicated in FIG. 3), which is supported by multiple lines of evidence, such as STS and BAC map data.

The nucleic acid molecules are also useful in making vectors containing the gene regulatory regions of the nucleic acid molecules of the present invention.

The nucleic acid molecules are also useful for designing ribozymes corresponding to all, or a part, of the mRNA produced from the nucleic acid molecules described herein.

The nucleic acid molecules are also useful for making vectors that express part, or all, of the peptides.

The nucleic acid molecules are also useful for constructing host cells expressing a part, or all, of the nucleic acid molecules and peptides.

The nucleic acid molecules are also useful for constructing transgenic animals expressing all, or a part, of the nucleic acid molecules and peptides.

The nucleic acid molecules are also useful as hybridization probes for determining the presence, level, form and distribution of nucleic acid expression. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain. Accordingly, the probes can be used to detect the presence of, or to determine levels of, a specific nucleic acid molecule in cells, tissues, and in organisms. The nucleic acid whose level is determined can be DNA or RNA. Accordingly, probes corresponding to the peptides described herein can be used to assess expression and/or gene copy number in a given cell, tissue, or organism. These uses are relevant for diagnosis of disorders involving an increase or decrease in kinase protein expression relative to normal results.

In vitro techniques for detection of mRNA include Northern hybridizations and in situ hybridizations. In vitro techniques for detecting DNA includes Southern hybridizations and in situ hybridization.

Probes can be used as a part of a diagnostic test kit for identifying cells or tissues that express a kinase protein, such as by measuring a level of a kinase-encoding nucleic acid in a sample of cells from a subject e.g., mRNA or genomic DNA, or determining if a kinase gene has been mutated. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain.

Nucleic acid expression assays are useful for drug screening to identify compounds that modulate kinase nucleic acid expression.

The invention thus provides a method for identifying a compound that can be used to treat a disorder associated with nucleic acid expression of the kinase gene, particularly biological and pathological processes that are mediated by the kinase in cells and tissues that express it. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland. The method typically includes assaying the ability of the compound to modulate the expression of the kinase nucleic acid and thus identifying a compound that can be used to treat a disorder characterized by undesired kinase nucleic acid expression. The assays can be performed in cell-based and cell-free systems. Cell-based assays include cells naturally expressing the kinase nucleic acid or recombinant cells genetically engineered to express specific nucleic acid sequences.

The assay for kinase nucleic acid expression can involve direct assay of nucleic acid levels, such as mRNA levels, or on collateral compounds involved in the signal pathway. Further, the expression of genes that are up- or down-regulated in response to the kinase protein signal pathway can also be assayed. In this embodiment the regulatory regions of these genes can be operably linked to a reporter gene such as luciferase.

Thus, modulators of kinase gene expression can be identified in a method wherein a cell is contacted with a candidate compound and the expression of mRNA determined. The level of expression of kinase mRNA in the presence of the candidate compound is compared to the level of expression of kinase mRNA in the absence of the candidate compound. The candidate compound can then be identified as a modulator of nucleic acid expression based on this comparison and be used, for example to treat a disorder characterized by aberrant nucleic acid expression. When expression of mRNA is statistically significantly greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of nucleic acid expression. When nucleic acid expression is statistically significantly less in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of nucleic acid expression.

The invention further provides methods of treatment, with the nucleic acid as a target, using a compound identified through drug screening as a gene modulator to modulate kinase nucleic acid expression in cells and tissues that express the kinase. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain. Modulation includes both up-regulation (i.e. activation or agonization) or down-regulation (suppression or antagonization) or nucleic acid expression.

Alternatively, a modulator for kinase nucleic acid expression can be a small molecule or drug identified using the screening assays described herein as long as the drug or small molecule inhibits the kinase nucleic acid expression in the cells and tissues that express the protein. Experimental data as provided in FIG. 1 indicates expression in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant and fetal brain, and thyroid gland.

The nucleic acid molecules are also useful for monitoring the effectiveness of modulating compounds on the expression or activity of the kinase gene in clinical trials or in a treatment regimen. Thus, the gene expression pattern can serve as a barometer for the continuing effectiveness of treatment with the compound, particularly with compounds to which a patient can develop resistance. The gene expression pattern can also serve as a marker indicative of a physiological response of the affected cells to the compound. Accordingly, such monitoring would allow either increased administration of the compound or the administration of alternative compounds to which the patient has not become resistant. Similarly, if the level of nucleic acid expression falls below a desirable level, administration of the compound could be commensurately decreased.

The nucleic acid molecules are also useful in diagnostic assays for qualitative changes in kinase nucleic acid expression, and particularly in qualitative changes that lead to pathology. The nucleic acid molecules can be used to detect mutations in kinase genes and gene expression products such as mRNA. The nucleic acid molecules can be used as hybridization probes to detect naturally occurring genetic mutations in the kinase gene and thereby to determine whether a subject with the mutation is at risk for a disorder caused by the ,mutation. Mutations include deletion, addition, or substitution of one or more nucleotides in the gene, chromosomal rearrangement, such as inversion or transposition, modification of genomic DNA, such as aberrant methylation patterns or changes in gene copy number, such as amplification. Detection of a mutated form of the kinase gene associated with a dysfunction provides a diagnostic tool for an active disease or susceptibility to disease when the disease results from overexpression, underexpression, or altered expression of a kinase protein.

Individuals carrying mutations in the kinase gene can be detected at the nucleic acid level by a variety of techniques. FIG. 3 provides information on SNPs that have been found in the gene encoding the kinase protein of the present invention. SNPs were identified at 42 different nucleotide positions. Some of these SNPs, which are located outside the ORF and in introns, may affect gene transcription. The gene encoding the novel kinase protein of the present invention is located on a genome component that has been mapped to human chromosome 22 (as indicated in FIG. 3), which is supported by multiple lines of evidence, such as STS and BAC map data. Genomic DNA can be analyzed directly or can be amplified by using PCR prior to analysis. RNA or cDNA can be used in the same way. In some uses, detection of the mutation involves the use of a probe/primer in a polymerase chain reaction (PCR) (see, e.g. U.S. Pat. Nos. 4,683,195 and 4,683,202), such as anchor PCR or RACE PCR, or, alternatively, in a ligation chain reaction (LCR) (see, e.g., Landegran et al., *Science* 241:1077–1080 (1988); and Nakazawa et al., *PNAS* 91:360–364 (1994)), the latter of which can be particularly useful for detecting point mutations in the gene (see Abravaya et al., *Nucleic Acids Res.* 23:675–682 (1995)). This method can include the steps of collecting a sample of cells from a patient, isolating nucleic acid (e.g., genomic, mRNA or both) from the cells of the sample, contacting the nucleic acid sample with one or more primers which specifically hybridize to a gene under conditions such that hybridization and amplification of the gene (if present) occurs, and detecting the presence or absence of an amplification product, or detecting the size of the amplification product and comparing the length to a control sample. Deletions and insertions can be detected by a change in size of the amplified product compared to the normal genotype. Point mutations can be identified by hybridizing amplified DNA to normal RNA or antisense DNA sequences.

Alternatively, mutations in a kinase gene can be directly identified, for example, by alterations in restriction enzyme digestion patterns determined by gel electrophoresis.

Further, sequence-specific ribozymes (U.S. Pat. No. 5,498,531) can be used to score for the presence of specific mutations by development or loss of a ribozyme cleavage site. Perfectly matched sequences can be distinguished from mismatched sequences by nuclease cleavage digestion assays or by differences in melting temperature.

Sequence changes at specific locations can also be assessed by nuclease protection assays such as RNase and S1 protection or the chemical cleavage method. Furthermore, sequence differences between a mutant kinase gene and a wild-type gene can be determined by direct DNA sequencing. A variety of automated sequencing procedures can be utilized when performing the diagnostic assays (Naeve, C. W., (1995) Biotechniques 19:448), including sequencing by mass spectrometry (see, e.g., PCT International Publication No. WO 94/16101; Cohen et al., *Adv. Chromatogr.* 36:127–162 (1996); and Griffin et al., *Appl. Biochem. Biotechnol.* 38:147–159 (1993)).

Other methods for detecting mutations in the gene include methods in which protection from cleavage agents is used to detect mismatched bases in RNA/RNA or RNA/DNA duplexes (Myers etal., *Science* 230:1242 (1985)); Cotton et al., *PNAS* 85:4397 (1988); Saleeba et al., *Meth. Enzymol.* 21 7:286–295 (1992)), electrophoretic mobility of mutant and wild type nucleic acid is compared (Orita et al., *PNAS* 86:2766 (1989); Cotton et al., *Mutat. Res.* 285:125–144 (1993); and Hayashi et al., *Genet. Anal. Tech. Appl.* 9:73–79 (1992)), and movement of mutant or wild-type fragments in polyacrylamide gels containing a gradient of denaturant is assayed using denaturing gradient gel electrophoresis (Myers et al., *Nature* 313:495 (1985)). Examples of other techniques for detecting point mutations include selective oligonucleotide hybridization, selective amplification, and selective primer extension.

The nucleic acid molecules are also useful for testing an individual for a genotype that while not necessarily causing the disease, nevertheless affects the treatment modality. Thus, the nucleic acid molecules can be used to study the relationship between an individual's genotype and the individual's response to a compound used for treatment (pharmacogenomic relationship). Accordingly, the nucleic acid molecules described herein can be used to assess the mutation content of the kinase gene in an individual in order to select an appropriate compound or dosage regimen for treatment. FIG. 3 provides information on SNPs that have been found in the gene encoding the kinase protein of the present invention. SNPs were identified at 42 different nucleotide positions. Some of these SNPs, which are located outside the ORF and in introns, may affect gene transcription.

Thus nucleic acid molecules displaying genetic variations that affect treatment provide a diagnostic target that can be used to tailor treatment in an individual. Accordingly, the production of recombinant cells and animals containing these polymorphisms allow effective clinical design of treatment compounds and dosage regimens.

The nucleic acid molecules are thus useful as antisense constructs to control kinase gene expression in cells, tissues, and organisms. A DNA antisense nucleic acid molecule is designed to be complementary to a region of the gene involved in transcription, preventing transcription and hence production of kinase protein. An antisense RNA or DNA nucleic acid molecule would hybridize to the mRNA and thus block translation of mRNA into kinase protein.

Alternatively, a class of antisense molecules can be used to inactivate mRNA in order to decrease expression of kinase nucleic acid. Accordingly, these molecules can treat a disorder characterized by abnormal or undesired kinase nucleic acid expression. This technique involves cleavage by means of ribozymes containing nucleotide sequences complementary to one or more regions in the mRNA that attenuate the ability of the mRNA to be translated. Possible regions include coding regions and particularly coding regions corresponding to the catalytic and other functional activities of the kinase protein, such as substrate binding.

The nucleic acid molecules also provide vectors for gene therapy in patients containing cells that are aberrant in kinase gene expression. Thus, recombinant cells, which include the patient's cells that have been engineered ex vivo and returned to the patient, are introduced into an individual where the cells produce the desired kinase protein to treat the individual.

The invention also encompasses kits for detecting the presence of a kinase nucleic acid in a biological sample. Experimental data as provided in FIG. 1 indicates that the kinase proteins of the present invention are expressed in humans in teratocarcinoma, ovary, testis, nervous tissue, bladder, infant brain, and thyroid gland, as indicated by virtual northern blot analysis. In addition, PCR-based tissue screening panels indicate expression in fetal brain. For example, the kit can comprise reagents such as a labeled or labelable nucleic acid or agent capable of detecting kinase nucleic acid in a biological sample; means for determining the amount of kinase nucleic acid in the sample; and means for comparing the amount of kinase nucleic acid in the sample with a standard. The compound or agent can be packaged in a suitable container. The kit can further comprise instructions for using the kit to detect kinase protein nRNA or DNA.

Nucleic Acid Arrays

The present invention further provides nucleic acid detection kits, such as arrays or microarrays of nucleic acid molecules that are based on the sequence information provided in FIGS. 1 and 3 (SEQ ID NOS:1 and 3).

As used herein "Arrays" or "Microarrays" refers to an array of distinct polynucleotides or oligonucleotides synthesized on a substrate, such as paper, nylon or other type of membrane, filter, chip, glass slide, or any other suitable solid support. In one embodiment, the microarray is prepared and used according to the methods described in U.S. Pat. No. 5,837,832, Chee et al., PCT application W095/11995 (Chee et al.), Lockhart, D. J. et al. (1996; *Nat. Biotech.* 14: 1675–1680) and Schena, M. et al. (1996; *Proc. Natl. Acad. Sci.* 93: 10614–10619), all of which are incorporated herein in their entirety by reference. In other embodiments, such arrays are produced by the methods described by Brown et al., U.S. Pat. No. 5,807,522.

The microarray or detection kit is preferably composed of a large number of unique, single-stranded nucleic acid sequences, usually either synthetic antisense oligonucleotides or fragments of cDNAs, fixed to a solid support. The oligonucleotides are preferably about 6–60 nucleotides in length, more preferably 15–30 nucleotides in length, and most preferably about 20–25 nucleotides in length. For a certain type of microarray or detection kit, it may be preferable to use oligonucleotides that are only 7–20 nucleotides in length. The microarray or detection kit may contain oligonucleotides that cover the known 5', or 3', sequence, sequential oligonucleotides which cover the full length sequence; or unique oligonucleotides selected from particular areas along the length of the sequence. Polynucleotides used in the microarray or detection kit may be oligonucleotides that are specific to a gene or genes of interest.

In order to produce oligonucleotides to a known sequence for a microarray or detection kit, the gene(s) of interest (or an ORF identified from the contigs of the present invention) is typically examined using a computer algorithm which starts at the 5' or at the 3' end of the nucleotide sequence. Typical algorithms will then identify oligomers of defined length that are unique to the gene, have a GC content within a range suitable for hybridization, and lack predicted secondary structure that may interfere with hybridization. In certain situations it may be appropriate to use pairs of oligonucleotides on a microarray or detection kit. The "pairs" will be identical, except for one nucleotide that preferably is located in the center of the sequence. The second oligonucleotide in the pair (mismatched by one) serves as a control. The number of oligonucleotide pairs may range from two to one million. The oligomers are synthesized at designated areas on a substrate using a light-directed chemical process. The substrate may be paper, nylon or other type of membrane, filter, chip, glass slide or any other suitable solid support.

In another aspect, an oligonucleotide may be synthesized on the surface of the substrate by using a chemical coupling procedure and an ink jet application apparatus, as described in PCT application W095/251116 (Baldeschweiler et al.) which is incorporated herein in its entirety by reference. In another aspect, a "gridded" array analogous to a dot (or slot) blot may be used to arrange and link cDNA fragments or oligonucleotides to the surface of a substrate using a vacuum system, thermal, UV, mechanical or chemical bonding procedures. An array, such as those described above, may be produced by hand or by using available devices (slot blot or dot blot apparatus), materials (any suitable solid support), and machines (including robotic instruments), and may contain 8, 24, 96, 384, 1536, 6144 or more oligonucleotides, or any other number between two and one million which lends itself to the efficient use of commercially available instrumentation.

In order to conduct sample analysis using a microarray or detection kit, the RNA or DNA from a biological sample is made into hybridization probes. The mRNA is isolated, and cDNA is produced and used as a template to make antisense RNA (aRNA). The aRNA is amplified in the presence of fluorescent nucleotides, and labeled probes are incubated with the microarray or detection kit so that the probe sequences hybridize to complementary oligonucleotides of the microarray or detection kit. Incubation conditions are adjusted so that hybridization occurs with precise complementary matches or with various degrees of less complementarity. After removal of nonhybridized probes, a scanner is used to determine the levels and patterns of fluorescence. The scanned images are examined to determine degree of complementarity and the relative abundance of each oligonucleotide sequence on the microarray or detection kit. The biological samples may be obtained from any bodily fluids (such as blood, urine, saliva, phlegm, gastric juices, etc.), cultured cells, biopsies, or other tissue preparations. A detection system may be used to measure the absence, presence, and amount of hybridization for all of the distinct sequences simultaneously. This data may be used for large-scale correlation studies on the sequences, expression patterns, mutations, variants, or polymorphisms among samples.

Using such arrays, the present invention provides methods to identify the expression of the kinase proteins/peptides of the present invention. In detail, such methods comprise incubating a test sample with one or more nucleic acid molecules and assaying for binding of the nucleic acid molecule with components within the test sample. Such assays will typically involve arrays comprising many genes, at least one of which is a gene of the present invention and or alleles of the kinase gene of the present invention. FIG. 3 provides information on SNPs that have been found in the gene encoding the kinase protein of the present invention. SNPs were identified at 42 different nucleotide positions. Some of these SNPs, which are located outside the ORF and in introns, may affect gene transcription.

Conditions for incubating a nucleic acid molecule with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the nucleic acid molecule used in the assay. One skilled in the art will recognize that any one of the commonly available hybridization, amplification or array assay formats can readily be adapted to employ the novel fragments of the Human genome disclosed herein. Examples of such assays can be found in Chard, T, *An Introduction to Radioimmunoassay and Related Techniques*, Elsevier Science Publishers, Amsterdam, The Netherlands (1986); Bullock, G. R. et al., *Techniques in Immunocytochemistry*, Academic Press, Orlando, Fla. Vol. 1 (1 982), Vol. 2 (1983), Vol. 3 (1985); Tijssen, P., *Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology*, Elsevier Science Publishers, Amsterdam, The Netherlands (1985).

The test samples of the present invention include cells, protein or membrane extracts of cells. The test sample used in the above-described method will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods for preparing nucleic acid extracts or of cells are well known in the art and can be readily be adapted in order to obtain a sample that is compatible with the system utilized.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the assays of the present invention.

Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the nucleic acid molecules that can bind to a fragment of the Human genome disclosed herein; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of a bound nucleic acid.

In detail, a compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers, strips of plastic, glass or paper, or arraying material such as silica. Such containers allows one to efficiently transfer reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such containers will include a container which will accept the test sample, a container which contains the nucleic acid probe, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, etc.), and containers which contain the reagents used to detect the bound probe. One skilled in the art will readily recognize that the previously unidentified kinase gene of the present invention can be routinely identified using the sequence information disclosed herein can be readily incorporated into one of the established kit formats which are well known in the art, particularly expression arrays.

Vectors/host Cells

The invention also provides vectors containing the nucleic acid molecules described herein. The term "vector" refers to a vehicle, preferably a nucleic acid molecule, which can transport the nucleic acid molecules. When the vector is a nucleic acid molecule, the nucleic acid molecules are covalently linked to the vector nucleic acid. With this aspect of the invention, the vector includes a plasmid, single or double stranded phage, a single or double stranded RNA or DNA viral vector, or artificial chromosome, such as a BAC, PAC, YAC, OR MAC.

A vector can be maintained in the host cell as an extrachromosomal element where it replicates and produces additional copies of the nucleic acid molecules. Alternatively, the vector may integrate into the host cell genome and produce additional copies of the nucleic acid molecules when the host cell replicates.

The invention provides vectors for the maintenance (cloning vectors) or vectors for expression (expression vectors) of the nucleic acid molecules. The vectors can function in prokaryotic or eukaryotic cells or in both (shuttle vectors).

Expression vectors contain cis-acting regulatory regions that are operably linked in the vector to the nucleic acid molecules such that transcription of the nucleic acid molecules is allowed in a host cell. The nucleic acid molecules can be introduced into the host cell with a separate nucleic acid molecule capable of affecting transcription. Thus, the second nucleic acid molecule may provide a trans-acting factor interacting with the cis-regulatory control region to allow transcription of the nucleic acid molecules from the vector. Alternatively, a trans-acting factor may be supplied by the host cell. Finally, a trans-acting factor can be produced from the vector itself It is understood, however, that in some embodiments, transcription and/or translation of the nucleic acid molecules can occur in a cell-free system.

The regulatory sequence to which the nucleic acid molecules described herein can be operably linked include promoters for directing mRNA transcription. These include, but are not limited to, the left promoter from bacteriophage λ, the lac, TRP, and TAC promoters from *E. coli*, the early and late promoters from SV40, the CMV immediate early promoter, the adenovirus early and late promoters, and retrovirus long-terminal repeats.

In addition to control regions that promote transcription, expression vectors may also include regions that modulate transcription, such as repressor binding sites and enhancers. Examples include the SV40 enhancer, the cytomegalovirus immediate early enhancer, polyoma enhancer, adenovirus enhancers, and retrovirus LTR enhancers.

In addition to containing sites for transcription initiation and control, expression vectors can also contain sequences necessary for transcription termination and, in the transcribed region a ribosome binding site for translation. Other regulatory control elements for expression include initiation and termination codons as well as polyadenylation signals. The person of ordinary skill in the art would be aware of the numerous regulatory sequences that are useful in expression vectors. Such regulatory sequences are described, for example, in Sambrook et al., *Molecular Cloning: A Laboratory Manual.* 2nd. ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., (1989).

A variety of expression vectors can be used to express a nucleic acid molecule. Such vectors include chromosomal, episomal, and virus-derived vectors, for example vectors derived from bacterial plasmids, from bacteriophage, from yeast episomes, from yeast chromosomal elements, including yeast artificial chromosomes, from viruses such as baculoviruses, papovaviruses such as SV40, Vaccinia viruses, adenoviruses, poxviruses, pseudorabies viruses, and retroviruses. Vectors may also be derived from combinations of these sources such as those derived from plasmid and bacteriophage genetic elements, e.g. cosmids and phagemids. Appropriate cloning and expression vectors for prokaryotic and eukaryotic hosts are described in Sambrook et al., *Molecular Cloning: A Laboratory Manual.* 2nd. ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., (1989).

The regulatory sequence may provide constitutive expression in one or more host cells (i.e. tissue specific) or may provide for inducible expression in one or more cell types such as by temperature, nutrient additive, or exogenous factor such as a hormone or other ligand. A variety of vectors providing for constitutive and inducible expression in prokaryotic and eukaryotic hosts are well known to those of ordinary skill in the art.

The nucleic acid molecules can be inserted into the vector nucleic acid by well-known methodology. Generally, the DNA sequence that will ultimately be expressed is joined to an expression vector by cleaving the DNA sequence and the expression vector with one or more restriction enzymes and then ligating the fragments together. Procedures for restriction enzyme digestion and ligation are well known to those of ordinary skill in the art.

The vector containing the appropriate nucleic acid molecule can be introduced into an appropriate host cell for propagation or expression using well-known techniques. Bacterial cells include, but are not limited to, *E. coli*, Streptomyces, and *Salmonella typhimurium*. Eukaryotic cells include, but are not limited to, yeast, insect cells such as Drosophila, animal cells such as COS and CHO cells, and plant cells.

As described herein, it may be desirable to express the peptide as a fusion protein. Accordingly, the invention provides fusion vectors that allow for the production of the peptides. Fusion vectors can increase the expression of a recombinant protein, increase the solubility of the recombinant protein, and aid in the purification of the protein by acting for example as a ligand for affinity purification. A proteolytic cleavage site may be introduced at the junction of the fusion moiety so that the desired peptide can ultimately be separated from the fusion moiety. Proteolytic enzymes include, but are not limited to, factor Xa, thrombin, and enterokinase. Typical fusion expression vectors include pGEX (Smith et al., *Gene* 67:31–40 (1988)), pMAL (New England Biolabs, Beverly, Mass.) and pRIT5 (Pharmacia, Piscataway, N.J.) which fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein. Examples of suitable inducible non-fusion *E. coli* expression vectors include pTrc (Amann et al., *Gene* 69:301–315 (1988)) and pET 11 d (Studier et al., *Gene Expression Technology: Methods in Enzymology* 185:60–89 (1990)).

Recombinant protein expression can be maximized in host bacteria by providing a genetic background wherein the host cell has an impaired capacity to proteolytically cleave the recombinant protein. (Gottesman, S., *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, Calif. (1990) 119–128). Alternatively, the sequence of the nucleic acid molecule of interest can be altered to provide preferential codon usage for a specific host cell, for example *E. coli.* (Wada et al., *Nucleic Acids Res.* 20:2111–2118 (1992)).

The nucleic acid molecules can also be expressed by expression vectors that are operative in yeast. Examples of vectors for expression in yeast e.g., *S. cerevisiae* include pYepSec1 (Baldari, et al., *EMBO J.* 6:229–234 (1987)), pMFa (Kurjan et al., *Cell* 30:933–943(1982)), pJRY88 (Schultz et al., *Gene* 54:113–123 (1987)), and pYES2 (Invitrogen Corporation, San Diego, Calif.).

The nucleic acid molecules can also be expressed in insect cells using, for example, baculovirus expression vectors. Baculovirus vectors available for expression of proteins in cultured insect cells (e.g., Sf 9 cells) include the pAc series (Smith et al., *Mol. Cell Biol.* 3:2156–2165 (1983)) and the pVL series (Lucklow et al., *Virology* 170:31–39 (1989)).

In certain embodiments of the invention, the nucleic acid molecules described herein are expressed in mammalian cells using mammalian expression vectors. Examples of mammalian expression vectors include pCDM8 (Seed, B. *Nature* 329:840(1987)) and pMT2PC (Kaufman et al., *EMBO J.* 6:187–195 (1987)).

The expression vectors listed herein are provided by way of example only of the well-known vectors available to those of ordinary skill in the art that would be useful to express the nucleic acid molecules. The person of ordinary skill in the art would be aware of other vectors suitable for maintenance propagation or expression of the nucleic acid molecules described herein. These are found for example in Sambrook, J., Fritsh, E. F., and Maniatis, T. *Molecular Cloning: A Laboratory Manual.* 2nd, ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989.

The invention also encompasses vectors in which the nucleic acid sequences described herein are cloned into the vector in reverse orientation, but operably linked to a regulatory sequence that permits transcription of antisense RNA. Thus, an antisense transcript can be produced to all, or to a portion, of the nucleic acid molecule sequences described herein, including both coding and non-coding regions. Expression of this antisense RNA is subject to each of the parameters described above in relation to expression of the sense RNA (regulatory sequences, constitutive or inducible expression, tissue-specific expression).

The invention also relates to recombinant host cells containing the vectors described herein. Host cells therefore include prokaryotic cells, lower eukaryotic cells such as yeast, other eukaryotic cells such as insect cells, and higher eukaryotic cells such as mammalian cells.

The recombinant host cells are prepared by introducing the vector constructs described herein into the cells by techniques readily available to the person of ordinary skill in the art. These include, but are not limited to, calcium phosphate transfection, DEAE-dextran-mediated transfection, cationic lipid-mediated transfection, electroporation, transduction, infection, lipofection, and other techniques such as those found in Sambrook, et al. (*Molecular Cloning: A Laboratory Manual.* 2nd, ed, Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989).

Host cells can contain more than one vector. Thus, different nucleotide sequences can be introduced on different vectors of the same cell. Similarly, the nucleic acid molecules can be introduced either alone or with other nucleic acid molecules that are not related to the nucleic acid molecules such as those providing trans-acting factors for expression vectors. When more than one vector is introduced into a cell, the vectors can be introduced independently, co-introduced or joined to the nucleic acid molecule vector.

In the case of bacteriophage and viral vectors, these can be introduced into cells as packaged or encapsulated virus by standard procedures for infection and transduction. Viral vectors can be replication-competent or replication-defective. In the case in which viral replication is defective, replication will occur in host cells providing functions that complement the defects.

Vectors generally include selectable markers that enable the selection of the subpopulation of cells that contain the recombinant vector constructs. The marker can be contained in the same vector that contains the nucleic acid molecules described herein or may be on a separate vector. Markers include tetracycline or ampicillin-resistance genes for prokaryotic host cells and dihydrofolate reductase or neomycin resistance for eukaryotic host cells. However, any marker that provides selection for a phenotypic trait will be effective.

While the mature proteins can be produced in bacteria, yeast, mammalian cells, and other cells under the control of the appropriate regulatory sequences, cell-free transcription and translation systems can also be used to produce these proteins using RNA derived from the DNA constructs described herein.

Where secretion of the peptide is desired, which is difficult to achieve with multi-transmembrane domain containing proteins such as kinases, appropriate secretion signals are incorporated into the vector. The signal sequence can be endogenous to the peptides or heterologous to these peptides.

Where the peptide is not secreted into the medium, which is typically the case with kinases, the protein can be isolated from the host cell by standard disruption procedures, including freeze thaw, sonication, mechanical disruption, use of lysing agents and the like. The peptide can then be recovered and purified by well-known purification methods including ammonium sulfate precipitation, acid extraction, anion or cationic exchange chromatography, phosphocellulose chromatography, hydrophobic-interaction chromatography, affinity chromatography, hydroxylapatite chromatography, lectin chromatography, or high performance liquid chromatography.

It is also understood that depending upon the host cell in recombinant production of the peptides described herein, the peptides can have various glycosylation patterns, depending upon the cell, or maybe non-glycosylated as when produced in bacteria. In addition, the peptides may include an initial modified methionine in some cases as a result of a host-mediated process.

Uses of Vectors and Host Cells

The recombinant host cells expressing the peptides described herein have a variety of uses. First, the cells are useful for producing a kinase protein or peptide that can be further purified to produce desired amounts of kinase protein or fragments. Thus, host cells containing expression vectors are useful for peptide production.

Host cells are also useful for conducting cell-based assays involving the kinase protein or kinase protein fragments, such as those described above as well as other formats known in the art. Thus, a recombinant host cell expressing a native kinase protein is useful for assaying compounds that stimulate or inhibit kinase protein function.

Host cells are also useful for identifying kinase protein mutants in which these functions are affected. If the mutants naturally occur and give rise to a pathology, host cells containing the mutations are useful to assay compounds that have a desired effect on the mutant kinase protein (for example, stimulating or inhibiting function) which may not be indicated by their effect on the native kinase protein.

Genetically engineered host cells can be further used to produce non-human transgenic animals. A transgenic animal is preferably a mammal, for example a rodent, such as a rat or mouse, in which one or more of the cells of the animal include a transgene. A transgene is exogenous DNA which is integrated into the genome of a cell from which a transgenic animal develops and which remains in the genome of the mature animal in one or more cell types or tissues of the transgenic animal. These animals are useful for studying the function of a kinase protein and identifying and evaluating modulators of kinase protein activity. Other examples of transgenic animals include non-human primates, sheep, dogs, cows, goats, chickens, and amphibians.

A transgenic animal can be produced by introducing nucleic acid into the male pronuclei of a fertilized oocyte, e.g., by microinjection, retroviral infection, and allowing the oocyte to develop in a pseudopregnant female foster animal. Any of the kinase protein nucleotide sequences can be introduced as a transgene into the genome of a non-human animal, such as a mouse.

Any of the regulatory or other sequences useful in expression vectors can form part of the transgenic sequence. This includes intronic sequences and polyadenylation signals, if not already included. A tissue-specific regulatory sequence (s) can be operably linked to the transgene to direct expression of the kinase protein to particular cells.

Methods for generating transgenic animals via embryo manipulation and microinjection, particularly animals such as mice, have become conventional in the art and are described, for example, in U.S. Pat. Nos. 4,736,866 and 4,870,009, both by Leder et al, U.S. Pat. No. 4,873,191 by Wagner et al. and in Hogan, B., *Manipulating the Mouse Embryo*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986). Similar methods are used for production of other transgenic animals. A transgenic founder animal can be identified based upon the presence of the transgene in its genome and/or expression of transgenic mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene can further be bred to other transgenic animals carrying other transgenes. A transgenic animal also includes animals in which the entire animal or tissues in the animal have been produced using the homologously recombinant host cells described herein.

In another embodiment, transgenic non-human animals can be produced which contain selected systems that allow for regulated expression of the transgene. One example of such a system is the cre/loxP recombinase system of bacteriophage P1. For a description of the cre/loxP recombinase system, see, e.g., Lakso et al. *PNAS* 89:6232–6236 (1992). Another example of a recombinase system is the FLP recombinase system of *S. cerevisiae* (O'Gorman et al. *Science* 251:1351–1355 (1991). If a cre/loxP recombinase system is used to regulate expression of the transgene, animals containing transgenes encoding both the Cre recombinase and a selected protein is required. Such animals can be provided through the construction of "double" transgenic animals, e.g., by mating two transgenic animals, one containing a transgene encoding a selected protein and the other containing a transgene encoding a recombinase.

Clones of the non-human transgenic animals described herein can also be produced according to the methods described in Wilmut, I. et al. *Nature* 385:810–813 (1997) and PCT International Publication Nos. WO 97/07668 and WO 97/07669. In brief, a cell, e.g., a somatic cell, from the transgenic animal can be isolated and induced to exit the growth cycle and enter G_o phase. The quiescent cell can then be fused, e.g., through the use of electrical pulses, to an enucleated oocyte from an animal of the same species from which the quiescent cell is isolated. The reconstructed oocyte is then cultured such that it develops to morula or blastocyst and then transferred to pseudopregnant female foster animal. The offspring born of this female foster animal will be a clone of the animal from which the cell, e.g., the somatic cell, is isolated.

Transgenic animals containing recombinant cells that express the peptides described herein are useful to conduct the assays described herein in an in vivo context. Accordingly, the various physiological factors that are present in vivo and that could effect substrate binding, kinase protein activation, and signal transduction, may not be evident from in vitro cell-free or cell-based assays. Accordingly, it is useful to provide non-human transgenic animals to assay in vivo kinase protein function, including substrate interaction, the effect of specific mutant kinase proteins on kinase protein function and substrate interaction, and the effect of chimeric kinase proteins. It is also possible to assess the effect of null mutations, that is, mutations that substantially or completely eliminate one or more kinase protein functions.

All publications and patents mentioned in the above specification are herein incorporated by reference. Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments. Indeed, various modifications of the above-described modes for carrying out the invention which are obvious to those skilled in the field of molecular biology or related fields are intended to be within the scope of the following claims.

SEQUENCE LISTING

```
<160> NUMBER OF SEQ ID NOS: 4

<210> SEQ ID NO 1
<211> LENGTH: 2320
<212> TYPE: DNA
<213> ORGANISM: Human

<400> SEQUENCE: 1

cccagggcgc cgtaggcggt gcatcccgtt cgcgcctggg gctgtggtct tcccgcgcct      60

gaggcggcgg cggcaggagc tgaggggagt tgtagggaac tgaggggagc tgctgtgtcc     120

cccgcctcct cctccccatt tccgcgctcc cgggaccatg tccgcgctgg cgggtgaaga     180

tgtctggagg tgtccaggct gtggggacca cattgctcca agccagatat ggtacaggac     240
```

-continued

```
tgtcaacgaa acctggcacg gctcttgctt ccggtgaaag tgatgcgcag cctggaccac     300

cccaatgtgc tcaagttcat tggtgtgctg tacaaggata agaagctgaa cctgctgaca     360

gagtacattg aggggggcac actgaaggac tttctgcgca gtatggatcc gttcccctgg     420

cagcagaagg tcaggtttgc caaaggaatc gcctccggaa tggacaagac tgtggtggtg     480

gcagactttg ggctgtcacg gctcatagtg gaagagagga aaagggcccc catggagaag     540

gccaccacca agaaacgcac cttgcgcaag aacgaccgca agaagcgcta cacggtggtg     600

ggaaacccct actggatggc ccctgagatg ctgaacggaa agagctatga tgagacggtg     660

gatatcttct cctttgggat cgttctctgt gagatcattg ggcaggtgta tgcagatcct     720

gactgccttc cccgaacact ggactttggc ctcaacgtga agcttttctg ggagaagttt     780

gttcccacag attgtccccc ggccttcttc ccgctggccg ccatctgctg cagactggag     840

cctgagagca gaccagcatt ctcgaaattg gaggactcct ttgaggccct ctccctgtac     900

ctgggggagc tgggcatccc gctgcctgca gagctggagg agttggacca cactgtgagc     960

atgcagtacg gcctgacccg ggactcacct ccctagccct ggcccagccc cctgcagggg    1020

ggtgttctac agccagcatt gcccctctgt gccccattcc tgctgtgagc agggccgtcc    1080

gggcttcctg tggattggcg gaatgtttag aagcagaaca aaccattcct attacctccc    1140

caggaggcaa gtgggcgcag caccagggaa atgtatctcc acaggttctg ggcctagtt     1200

actgtctgta aatccaatac ttgcctgaaa gctgtgaaga agaaaaaaac ccctggcctt    1260

tgggccagga ggaatctgtt actcgaatcc acccaggaac tccctggcag tggattgtgg    1320

gaggctcttg cttacactaa tcagcgtgac ctggacctgc tgggcaggat cccagggtga    1380

acctgcctgt gaactctgaa gtcactagtc cagctgggtg caggaggact tcaagtgtgt    1440

ggacgaaaga aagactgatg gctcaaaggg tgtgaaaaag tcagtgatgc tccccctttc    1500

tactccagat cctgtccttc ctggagcaag gttgagggag taggttttga agagtccctt    1560

aatatgtggt ggaacaggcc aggagttaga gaaagggctg gcttctgttt acctgctcac    1620

tggctctagc cagcccaggg accacatcaa tgtgagagga agcctccacc tcatgttttc    1680

aaacttaata ctggagactg gctgagaact tacggacaac atcctttctg tctgaaacaa    1740

acagtcacaa gcacaggaag aggctggggg actagaaaga ggccctgccc tctagaaagc    1800

tcagatcttg gcttctgtta ctcatactcg ggtgggctcc ttagtcagat gcctaaaaca    1860

ttttgcctaa agctcgatgg gttctggagg acagtgtggc ttgtcacagg cctagagtct    1920

gagggagggg agtgggagtc tcagcaatct cttggtcttg gcttcatggc aaccactgct    1980

caccccttcaa catgcctggt ttaggcagca gcttgggctg ggaagaggtg gtggcagagt    2040

ctcaaagctg agatgctgag agagatagct ccctgagctg ggccatctga cttctacctc    2100

ccatgtttgc tctcccaact cattagctcc tgggcagcat cctcctgagc cacatgtgca    2160

ggtactggaa aacctccatc ttggctccca gagctctagg aactcttcat cacaactaga    2220

tttgcctctt ctaagtgtct atgagcttgc accatattta ataaattggg aatgggtttg    2280

gggtattaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa                          2320
```

```
<210> SEQ ID NO 2
<211> LENGTH: 255
<212> TYPE: PRT
<213> ORGANISM: Human

<400> SEQUENCE: 2

Met Val Gln Asp Cys Gln Arg Asn Leu Ala Arg Leu Leu Leu Pro Val
```

-continued

```
1               5                   10                  15

Lys Val Met Arg Ser Leu Asp His Pro Asn Val Leu Lys Phe Ile Gly
            20                  25                  30

Val Leu Tyr Lys Asp Lys Lys Leu Asn Leu Leu Thr Glu Tyr Ile Glu
        35                  40                  45

Gly Gly Thr Leu Lys Asp Phe Leu Arg Ser Met Asp Pro Phe Pro Trp
    50                  55                  60

Gln Gln Lys Val Arg Phe Ala Lys Gly Ile Ala Ser Gly Met Asp Lys
65                  70                  75                  80

Thr Val Val Val Ala Asp Phe Gly Leu Ser Arg Leu Ile Val Glu Glu
                85                  90                  95

Arg Lys Arg Ala Pro Met Glu Lys Ala Thr Thr Lys Lys Arg Thr Leu
            100                 105                 110

Arg Lys Asn Asp Arg Lys Lys Arg Tyr Thr Val Val Gly Asn Pro Tyr
        115                 120                 125

Trp Met Ala Pro Glu Met Leu Asn Gly Lys Ser Tyr Asp Glu Thr Val
    130                 135                 140

Asp Ile Phe Ser Phe Gly Ile Val Leu Cys Glu Ile Ile Gly Gln Val
145                 150                 155                 160

Tyr Ala Asp Pro Asp Cys Leu Pro Arg Thr Leu Asp Phe Gly Leu Asn
                165                 170                 175

Val Lys Leu Phe Trp Glu Lys Phe Val Pro Thr Asp Cys Pro Pro Ala
            180                 185                 190

Phe Phe Pro Leu Ala Ala Ile Cys Cys Arg Leu Glu Pro Glu Ser Arg
        195                 200                 205

Pro Ala Phe Ser Lys Leu Glu Asp Ser Phe Glu Ala Leu Ser Leu Tyr
    210                 215                 220

Leu Gly Glu Leu Gly Ile Pro Leu Pro Ala Glu Leu Glu Glu Leu Asp
225                 230                 235                 240

His Thr Val Ser Met Gln Tyr Gly Leu Thr Arg Asp Ser Pro Pro
                245                 250                 255
```

```
<210> SEQ ID NO 3
<211> LENGTH: 59065
<212> TYPE: DNA
<213> ORGANISM: Human

<400> SEQUENCE: 3

tcatccttgc gcaggggcca tgctaacctt ctgtgtctca gtccaatttt aatgtatgtg      60

ctgctgaagc gagagtacca gaggtttttt tgatggcagt gacttgaact tatttaaaag     120

ataaggagga gccagtgagg gagaggggtg ctgtaaagat aactaaaagt gcacttcttc     180

taagaagtaa gatggaatgg gatccagaac aggggtgtca taccgagtag cccagccttt     240

gttccgtgga cactggggag tctaacccag agctgagata gcttgcagtg tggatgagcc     300

agctgagtac agcagatagg gaaaagaagc caaaaatctg aagtagggct ggggtgaagg     360

acagggaagg gctagagaga catttggaaa gtgaaaccag gtggatatga gaggagagag     420

tagagggtct tgatttcggg tctttcatgc ttaacccaaa gcaggtacta aagtatgtgt     480

tgattgaatg tctttgggtt tctcaagact ggagaaagca gggcaagctc tggagggtat     540

ggcaataaca agttatcttg aatatcctca tggtggaaag tcctgatcct gtttgaattt     600

tggaaataga aatcattcag agccaagaga ttgaattgtt gagtaagtgg gtggtcaggt     660

tacagactta attttgggtt aaaaagtaaa aacaagaaac aaggtgtggc tctaaaataa     720
```

-continued

```
tgagatgtgc tgggggtggg gcatggcagc tcataaactg accctgaaag ctcttacatg    780

taagagttcc aaaaatattt ccaaaacttg gaagattcat ttggatgttt gtgttcatta    840

aaatctctca ctaattcatt gtcttgtcca ctgtccgtaa cccaacctgg gattggtttg    900

agtgagtctc tcagactttc tgccttggag tttgtgagag agatggcata ctctgtgacc    960

actgtcaccc taaaaccaaa aaggcccctc ttgacaagga gtctgaggat tttagaccca   1020

ggaagaatga gtgatgggca tatatatatc ctattactga ggcatgagaa gagtggaatg   1080

ggtgggttga ggtggtgttt taaggcctct tgccagcttg tttaactctt ctctggggaa   1140

cgagggggac aactgtgtac attggctgct ccagaatgat gttgagcaat cttgaagtgc   1200

caggagctgt gctttgtcta ttcatggccc ctgtgcctgt gaaacagggt tcggtgactg   1260

tcactgtgcc tgtggcagtc tgtagttacc cagagagaac aaagctgcat acacagagcg   1320

cacaagggag tcttgtaaca accttgtcct gctttctagg gctgagtcag gtaccacagc   1380

ttgatctcag ctgtcctctt tatttcaaga agttgacatc tgagccatac caggagtatt   1440

gtattttgtt tgaggcctct ctttttggag gaacatggac cgactctgtg cttttgtcta   1500

tgctggtctc tgagctcaca caacccttca ccctcctttc tcagccagtg ataggtaagt   1560

cttccctatc ttgcaaggct cagctcaagt gtcagcttcc tctacaaaga ctttcctggt   1620

tcccctcatt ggagtgaaca agagttgaca tggtagaatg gaaagagcag aagctttaga   1680

atgagccaga cctgagtatg aatgctagat ccaccactta gctagtcaac cctgcccccct   1740

gcctcaagtt ttaattttcc tatccattaa gtgaatataa taatacctgt gtcacaggat   1800

tattttgaga attaaatgag attaggtcta tgaaagcacc tagcagagtt cttggcatat   1860

aggaggcatt cattaaatat ttgttcttcc cctttatac ccattacttt tctttttctg   1920

aactaaaata atacttggtt ctatctctga aataacatcc aagtgaaaaa tcaacaacat   1980

gaaagagcag ttcttttcca gtggatttgc ttcttaagga gcagagatta tgtaatctaa   2040

cagcctccaa catacaaaga gctttgtatc tagaacaggg gtccccagcc cctggaccgc   2100

caactggtac gggtctgtag cctgttagga accaggctgc acagcaggag gtgagcggcg   2160

ggccagtgag cattgctgcc tgagctctgc ctcctgtcag atcagtggtg gcattagatt   2220

ctcataggag tgtgaacccct attgtgaact gcacatgcaa gggatctggg ttgcatgctc   2280

cttatgagaa tctcactaat ggctgatgat ctgagttgga acagtttgat accaaaacca   2340

tccccccgcc ccccaacccc cagcctaggg tccgtggaaa aattggcccc tggtgccaaa   2400

aaggttgagg actgctgatc tagaggacca atttattcaa tgttggttga gtaaatgagc   2460

tcttggatta ggtgatggaa aaatctgaaa aaacagggct tttgaggaat aggaaaaggc   2520

agtaacatgt ttaacccaga gagaagtttc tggctgttgg ctgggaatag tcataggaag   2580

ggctgacact gaaaagaagg agattgtgtt cgtttcttct tctcagagct ataagcaaag   2640

gctgaaagtt ctagaaaaag gcaagttttg tttcagtaga aaaaaggata atcagaacca   2700

tttttagaaa atggaatgag actacttttg aggccatgag ttccttgtcc ctggagagat   2760

gagcagaggt tggacaagtg cttaccagag atcttgtgga ggcagaaact gtgcatctag   2820

cagagcattg gcctaaccct ttcaaatgag atgctgttaa ctcagtctta ttctacatgg   2880

taggaatcct gtccctttgc ctcctgctac tttgggcctc tcaacctctt ggttttgtgt   2940

gcaggtgaag atgtctggag gtgtccaggc tgtggggacc acattgctcc aagccagata   3000

tggtacagga ctgtcaacga aacctggcac ggctcttgct tccggtaggt gggcctatcc   3060

tcccatcttt accagtgtac tatgggccaa gcactatttc atgttctgat ggaaaacaca   3120
```

-continued

```
gaaacaagct tctgagttga gaatttcaat cttagggtgg ggaaaggaat gtaccaagga   3180
agagctcatg accaaacctc aagtgtggcc cccctgaacc caggttaaat tggaagagcc   3240
ataaatgggc cagctggagg cagggtgggg ggatgagagg agccctttcc agggttgtcc   3300
catatccctc actttatggg tgaggaaact gaggcccagg aagagtgact ttcctgtggc   3360
tgcactacag attatgcagg tacttcaaga gttgtttgta ttcttatttt attttatttt   3420
attttatttt attttatttt attttatgag agggattctt gctgttgccc aggctggagt   3480
gcagtggtgc aatctcggct cactgcaatc tctgcctgct gggttcaagt gatttttctg   3540
ccttagcttc ctgagtagct gagatgacag gcacctgcca ccatgcgcag ctaatttttg   3600
tattttagtg gagacggggg tttcaacatg ttggtcaggc tggtcttgaa ctcctgacct   3660
caaatgatgc acccacctcg acctcccaaa gtgctggaat tacaggcgtg aaccactgtg   3720
cccagccaag agttgttttt agtgtggttg gcagagccag ctcttccttc accacaggat   3780
gcctccctag gttcctactt tttgttacta gcttttatta tagctatatt attattatta   3840
ttattattat tattattatt attattgaga cagagtctcg ctctgtcgcc caggctggtg   3900
tacagtggtg cgatcccggg ctcactgcaa cctctgcctc ccgagttcaa gcagttctcc   3960
tgcctcagcc ccccgagtag gtgggactac aggcgcctgc caccacaccc ggctaatttt   4020
tgtattttta gtagagacgg ggtttcacct tgttgaccag gctggtctgg agctcctgac   4080
ctcaggtaag tgctagaatc acaggcgtga accactgcgc ccagccaaga gttgttttta   4140
gtgtggttgg cagagccagc tcttcctcac cacaggttgc ctccctaggt tcctactttt   4200
tgttactagc tttattatag ctacattatt attattattg ttattattat tgagacagag   4260
tctcgctctg tcgcccaggc tggtgtacag tgatgtgatc ttggctcact gcaacctctg   4320
cccccgagt tcaagcaatt ctcctgcttc agccccccta gtaggtggga ctccaggcac   4380
ctgccaccac gcccagctaa tttttgtatt tttagtagag gcggggtttc accttgttgg   4440
ccaggctggt ctcaaactcc tgacctcagg tgatccgcct gcctcggcct cccaaaatgt   4500
tgggattaca ggcatgagcc accgcgccct gcctatagct acattatttt tgtaggcagc   4560
tcagtttctt aaaaattata cagacttcaa atcagatttg ttcctgctgt ctgaggctca   4620
gtttcttcat ctggaaaatg gatggtaata atcttgttga gattgaatga aataatatat   4680
gcagtgtatc cagtacatgg tagacaccca gtgaatggtt attccttcct cccatcggat   4740
tggaattctc aagggtggga acttgtcttt atattcttca caacgtaaaa tagttgaaat   4800
ttgttggtgg aaagaagagc agtccactcc agaggctgga tgggcatgcc tggcccccaa   4860
ggtctgaagt ggtagggctg tgcctatatc ctgagaatga gatagactag gcaggcacct   4920
tgtgctgtag attccagctc ctgcacatag ctcttgttgt aaaacatccc tgtgcttata   4980
ccaagtaatt gagttgacct ttaaacactt gcctcttccc tgggaaccat ataggggatt   5040
ggcctggaga cgtctggcct ctggaagagt tggaaagcag ccatcattat tatcctttcc   5100
tttcagctat aactcagagc tctcaagtct tttctgtgga tcttattgcc ttggttcttg   5160
ccccttttac tcccagggaa gttgattctg tcttttctgt tccatttagt atgacaggag   5220
cagagaatgt cagagctgta agggacctta tagttaaagc ctttggctgg tcctttcatt   5280
ttatagctgg gactaataag taacgtcaaa acccaatgag ttcacagatt gggtctcgcc   5340
ttggcatgta acccatatgt tcatattctt gctgttttcc tatgtgtatg aatattttct   5400
atccaaaata agcaggacag ggtagagcaa gttaatcttt ggaatttctg gattctctta   5460
```

-continued

```
gagctaaaa acttcagaac tagaagaaac cacccactat atggtataac ccattcatat   5520
cacagatgag gcctgaaacc aaaaagactt gctcaggcca tggatgacaa gagctggccc   5580
tagcactgaa ctcttgggtc atttgtaggt ctagtcagat gctagcttgt tagctctgtg   5640
cgtgcgtgtg tgtgtgtgtg tgtgtgtgtg tgtgtgagat agagacagaa agataacata   5700
tgtacacaaa tacataaaga ggaagtagac acgttagcat ggtagataag agtacaggca   5760
ggccaggcgt ggtggctcac gcctgtaatc ccagcacttt gggaggccaa ggcaggtgga   5820
tcacctgagg tcaggaattc gagaccagcc tgaccaacat ggtgaaaccc catctctact   5880
aaatacagaa aaaaattagc ttggcatggt ggcacatgcc tgtaatccca gctacttggg   5940
aagctgaagc aggagaatcg cttgaatccg ggaagcagaa gttgcagtga gccgagattg   6000
tgccattaca gtctagcctg ggcaacaaga gggaaactcc atcgcaaaaa aacaaccacc   6060
accaagagta caggctatgg aatgagacta tggttttaaa tcctggcttt gcaatttatt   6120
aactagcctt aagtgacttc cctgagcttc aggcaccaat ctgtaaaatg aggataagaa   6180
tattactcat gccacatggt tgttagggag gattaaatgt gataacctat ataaagtggc   6240
tagcatagca tctgacatat agaaaactct taatagggcc ggacgtggtg gcttatgcct   6300
gtaatcctag cactctggga ggccgaggca gaaggatcgc ttgagcccat gagcccagga   6360
gtttgagacc agcctggcca acatggcaaa actccacctc tacaaaaaat acaaaaatat   6420
tagccaggcg tgatggcaca cacctgtagt cccagctact tgggaagctg aggagcgatg   6480
attacctgag cccagggata tcaaggctgt agtgagctgt gatcatgcca ctgtactcca   6540
tccagctggg ggacagagtg aaacccctgt ctcaaaacaa aacaaatgaa aaaaaaaacc   6600
cttaataatc agtaactgtc actttatatt atgttgtgag tgtgtgtcta tatacaccta   6660
tatgtataca tttctcttat tacacattca ttggtgatct gatgtggagc cccagggatt   6720
aagggcaact ttgaactacc ctgacacaat caagccaaat atcattcccg tggaggaagt   6780
agagtatcta ggttctgtct cctagttgca gctttacctt gaggacagag actctaatcc   6840
agctgtgctg aaggagcaca tctcctgact tctgagcttt cccctggtaa attcaaactg   6900
gatgtcacgg cgccctcaga tagagcctgg taatttgccc tggggagagt gactgtcttt   6960
tggatctaat ttgacttttg ccccagttgg aggaaaatct tcagggctag gaaggattgt   7020
atttgtctga ccccagagat aacctgggtt ttgaggaaca tggggcatca acctgaatgg   7080
tcttgtaaga tctctcccac gccagcttgc cagtgtttct ctgatgaatt tagagtacct   7140
gagtagtgca ggcctgctgg gaggaggact ctccctctgt gctactcaga gaaattcatt   7200
cttcaaggcc cccttccagc cttgctctta cccagctggg ctacagttac aataaaggaa   7260
atgacttttc ttctcccctt cccccagtac ctttgttttc ctagtcacag ggtggggctg   7320
gatattgaat ggagaaattg ctggggtcca tcctaaactc ctcccctcat ctctccctta   7380
cattacccca ttcttctgtc tgcagccaca tccataatcc tgcctctgtt agccttccga   7440
cagaccctca ggtgcccagg acaacaggaa gctacttaaa gctggaacct cagactgtgc   7500
aatggaggcc agtgacaaaa ctgaaagtag ctctgtcagt aattgtgctg gtgcgattag   7560
gcagctggcc agaatctttt ggatctcctg gacatatggc tgactagtcc tcccaagcct   7620
tcccaacagg cctctttttt ttcctttttt tcttttcttt tttttctttc tttctttctt   7680
tctttttttt ttttttttag gctagtgaag tgaaattgtg ggagtggaaa aggaacaaag   7740
aaatcggtaa ctggtagtga tcaattactt gtaaacacta ttgtacttgg accagcccag   7800
taggcctttt ttaaaactct gagttacctc tctttccttt ccttgagcag tgccattaat   7860
```

-continued

```
tctgtatctg gggcaatcct ttctgatgtt ctctggacct ggctctctct ccttaggaga   7920
ggccaggaga gtagccagag agcatgtcat ttgtagctga ggttaaagtg tggagctatc   7980
aatggtgacc tggcctcttg gcatgttagc aagccagagg accttgacaa ctttttgat   8040
gattgtccgt tcaccctgat caaaggtgtt tggcttagga ggagggaaga aaagctaccc   8100
ctattagtct tgatggcccc agcgtgggtc tctattgctt gacctggttc ctagcagcat   8160
tatcagaagg aaaatccacc gctcttaagg ctcctgggaa ctttcaggac ttcctttctc   8220
aggattgcaa acataagact atttgagctt tcacttttga aaagcggtta ctaataccta   8280
tactctggga aagggctaat gcagatagaa gactgtggtc actgcatcag gcaacagacc   8340
atttccgcta aatttagtga ctccaggaag gccagtgaag aaataacaca cgtagcaacc   8400
agagactgtg ttgtaatatg ttggctgaca gcagggtact ttctgtgatg ctgaaagcca   8460
cattcatttt ctctcccctc atccccatct aagcaagcct ggtagaatca taattacagt   8520
aataggtacc acttattgag tactctgtgc cagacaccct cctgagcata cgacatgcat   8580
agcacattta atccttacaa tgacttaata aaatgtagta ctagtcttac ctacttcgag   8640
aatagggaaa tggaggttac ttgtttaaag tcacagagct aataggtagc atagctgaga   8700
tttgaactca ggcattctta ctccttgcct gcaagagtct cttggcattc ttgaatgcaa   8760
gcatatttct taacctcact gaggctcagt ttcctcttat ataatatggg gtaaagagcc   8820
ctcaccctgc ctgccacaca ctggtagtgt cagataacat tgaagggtgt tagtttaaag   8880
gcttcatgga ctctataatg tcaacaaaag tgctgttaac tttcttctgg gtctcaggct   8940
cctgatgtag agtcagtgga gcaaccctgc catctgctgt tatgctgttg atgttgctgc   9000
cacacttact aacctaaacc tttgattctg gctgtggcct tctccagaag gtgtttactc   9060
atttgtccag tttatctttt aggaaacagc cagcccgtag atcattaagg ctggctattg   9120
gacagggggc tggggcctgc ctgacagagg aaggaagggc agacatctgg ttcttcctct   9180
gcccctacaa gagactccag cctgaccaca gagtggtact cctaggatgt agcagcagca   9240
tatgagcttg aatgtgcctt aatcctgctc tttactttga gaagagagaa ctaaggaccc   9300
acagatgttt cacagcttct ataggaggca gaggtagaaa aatggagaga gatgaggcca   9360
gagatagata actgatatta attaaacgtt gtattaagaa cctcacttag attatctgat   9420
tcaatcttca taataaccct gcaaccccca cctttttttg agaacagggt cttgctctgt   9480
tgtccaggct acagtgcact ggtacaatca tagttcactg cagtgtcaac ctcctgagct   9540
caagcaatcc tcccacctca gccttgcaag cagcttggac tacaggcgtg ccaccacacc   9600
ttgccatttt tttttatttt aagtagaaac aaggtcttat taatactatg ttgcccaggc   9660
tggtcttgaa ctccagcgat cctcctgccc cagcctccca aagtgcttgg gattacggaa   9720
gtaagccact gtgcctggcc agtgcaaccc ccattttata ctaaaacagg aaggcccaga   9780
aaggtttgga gtaacttgtc cagggtcaca cagatgatat ttgaactcag gtctccctgg   9840
ctcccaagag agtctgcttt ccactaggac tcccaggaga aaaaaaaaaa aaaaaacagt   9900
agacttggag acagaaaatc tgatttgagt cttagttgag ctaggctaac tgtgtaactg   9960
tgggcaagtt ccttagcccc tgtgagcctc agtttcttat ctgtaaaatg tcataaaaga  10020
aatccatctc atggagtagt tgtgatgatc aaggactctg aaaacattag aatggtttaa  10080
tgtgaaggat tagcagcagc acatggcaac attgtgcatc ttatattaac tatccaaata  10140
tatcaagcgt catttgctat atataaaagt catcaaatta ggcactgtgg gggatacgga  10200
```

-continued

```
gttggcatac tagcctggcc tcttaattaa ttcattaatt agcttattta tttttgagat  10260
aggtcttgct ctattgccca ggctggagtg cagtggcatg atgatagctt actatagcct  10320
caatctccca ggcttaaaca atcctcctga gtagctggga ctacaggcac acactaccat  10380
gcccagctaa ttttttttta atttttgta gagacagggt cttgctctgt tgcccaggct  10440
ggtctcaaac tcctgggctc gagatcctcc cacctgggcc tcacaaagtg ttgggattac  10500
aggtatgagc cacggcacct ggcctggtct cttaactggt tccctaagac agctggaaat  10560
agagaatgtc atggagcatt cctaaccatg ggctccagcc tggctttcat tctgtttctc  10620
ccctgaaaca acattccttt agtaatattc cgaataacag cttcatcagt ctgtctaccg  10680
accactcttc aggcttcatc ttatatgacc tcccaaactg cactaagggt tgtattagag  10740
aaaagtggat aaagttcgga gtcaggctgc ttgagcttaa atgccagctt cacttaccag  10800
ccacctgacc atgagtcagc tgcttaacca ttctttgcca cagtttcctt gtctatgaaa  10860
agggaaatgg ctcccacctc aaaaagttgt taacattaaa ttcaatcatg tattcaaagt  10920
cctgagcaga atgtctggcc atgactggga cttaacagat gttagcattt attattagta  10980
tctgtcagtc ttgaaatgtt ctcttccctt ggctttcatg acattccaca ctctcctggt  11040
tttctcttac ctctctggta atacctgttt gcttatcctt ctttgtccag ctctgggatg  11100
ttaccattcc ttcaggcgtg ctgttttctc cttaggcagt cttacacaca ctcatgactt  11160
ccttccattg tcctccacac actgatgacc ctaaaatcag tatctccagc ctaaaccttt  11220
ccactgagtt ctagacccat atgttgtact atcaacctgg cttgtccatt tgaatgtctt  11280
ccaggcactt cagactctct tctctagact ttgctggact ttcactcttc cccctaaaac  11340
tggctcctct tccactgaaa catgtatgtc attgagaggc accaccatcc acccagtgcc  11400
taagccagaa acctaggaat ccttgatacc tgttctctct catcctgcat atccaagcct  11460
atcagtttta tctctaaatt atattttggt aggtttactt ctttcctttt ctcccaccac  11520
caccctgctc caagctacca tcatctcacc tggatgtctg caatagcctc atctcccaca  11580
gccactctgc accccctaat ctgttctcta tagagcagtt ggaaggagtg atttttgttg  11640
tttgttttgt tttgttttag acagagtctc actctgttcc ccaaggctgg agtgcagtgg  11700
cacaatttcg gctcactgca acttctgcct cccgggttta agcaattctc ctgcctcagc  11760
ctcccaagta gctgggatta aggcaccggc ccccataccc agctaatttt tatattttta  11820
gtagagatgg ggttttgcca tgttggccaa gctagtctcg aactcctgac ctcaagtgat  11880
ccacctgcct cggcctccca aagtgctggg attacaggtg tgagccactg cacctggctg  11940
gaaggagtga tcttaaaaaa aaaaaaaaca aaaaaaaact tgactgtgtc actctgtgtt  12000
gtctctccta ccttgtatac ttccacaact tcccagtgtt cttggataaa gaccaaaatc  12060
cttaacttgg ccaggcgcgg tggctcacac ctatcatctc agcactttgg gaggccgagg  12120
caggcagatc atgaagtcaa gagattgaga ccatcctggc caacatggtg aaaccccatc  12180
tctactaaaa atacaaaaat tagctggtcg tggtggcgtg tgcctgtagt cccagctact  12240
tgggaggctg aggcaggaga atcacttgaa cctgggaggc agaggttgca gtgagcccag  12300
atcacgccac tgcactccag cctggtgaca gagtaagact ccatctcaaa aaaaaaaaaa  12360
aaaaaaaaaa ttccttaatt tggcctacag tagagccctc cgtaatgtgg cctctctcca  12420
catctccaca acctcctgct ccctgcactt cagcctcacc tctcttctgg acaggccctc  12480
cttctgacaa gggctttgtt cattctgctc cctctgccta gaatgccccc ttactctgtt  12540
cacttaactc ctgcttatcg tttagatctt tacctggatg gctcagagaa atatagaagt  12600
```

-continued

```
aattcctcac cctgaaaaat aggttaggtc cctgttttat gttttcatag acctttcctt  12660
tgaggctttt tttaaaaaag tagttttaat ctcacattta ttcatgtgat catctcctta  12720
atgatatctt aagacctcta atagaacaat ttggtcatgg actgtggggt ttttgcccct  12780
cattgtgtca gcactgagca tattgttggc ataggaggga tatttgttga atgaattgct  12840
agaggtggcc aagagatatg atgtaagtca ggcttttccc tgcccttccc cttccccttc  12900
cccacatcct tcctatagca gccaccgtgg ctgcagttac tgtaaatggc aagacggaat  12960
cagttccgga cattgggttg ttttagaaaa ttgcctgcaa gtgtcagggt gataagttaa  13020
agctttgtct tttgccctca gaggagctat cccatagtga gtagaagcca gagaagctga  13080
ccccaggagt ccttctttcc agcagcaggt cttgagctgc acttctctgt agctacaatc  13140
caggcaggaa caagccctag gtacctccgg agaggagggc aagagaggaa gaatgagttc  13200
agctactcta gccaccaaac tgattatgaa ttgccctgaa atctgaaaaa tttcaattcc  13260
aatcgtaagt ttgttttgtt tcattttgtt ttcttaaatt gtatatttga aagatggcat  13320
taactaaaga tatatattca atatagagtg gaaaaaatgg aatacttgca tagtatcttt  13380
tacttatagg tgatttatga tggggagtgg ggtggatagg ttggcagttc ccccaagaag  13440
ttggaaatga agtttgtcct ctgtgagttg aactaattag atccacaagt aatgaaagca  13500
gtattgtgtt gtagttaaga gcacactcta gaaccagatt gcttagtttc aaatcctggt  13560
tctgcctttt attatctgtg tactttgggc aagttacttg ccctttgtgt gcttcatttt  13620
tctcatctag aaaatggaga ggccaggcgt agtggctcat gcctataatc ccagcacttt  13680
gggaggccga ggcgggcaga tcacctgagg tgagaagttc aagaccagcc tggccaacat  13740
ggtgaaaccc tgtctctaca aaaatacaaa aattagccag gcatgatggc gggtgcctgt  13800
aatcccagct acccaggagc ctgaggcggg agaaacactt gaacctggaa ggcagaggtt  13860
gtagtgagcc aggattgcac cactgcactc cagcctgggt gacaagagct agactcagtc  13920
taaaaaaaaa aaaaaaaaac aaactggaga tacaggctgg gtgcagggct tacacttata  13980
atatcagcac tttgggaggc ctaggcggga ggattgcttg aactcaggag tttcaagatc  14040
agtctgggta acagagcaag acctcatccc cacaaaaaat caaaaattta gccaggcatg  14100
gtggctcatg cctgtggtcc cagctactca ggaggctgag gcgagaggat tgcttgagcc  14160
caggaggttg aggctgcagt gaaccatgac tgcaccacta catgccagcc tggatgacag  14220
agcaagaccc tatctcaaaa aaaaaaaaaa aaagaaacga gccaggcgcg tttgctcacg  14280
ccagtaatcc cagcactttg ggaggccaag gcaggtggat cacttgaggt caggagatcg  14340
agactagcct ggccaacatg gtgaaacccc atctcaactg aaaatacaaa aattagccag  14400
gcatggtggc atgctcctgt agtcccagct actcacttgg aggctgaggc acgagaatcg  14460
cttgaaccca ggaggcggag gttgcagtgg gccaacatca tgtcactgca ctccagcctg  14520
ggagacagag cgagactctg tctcaataaa taaataaaca taaaataaaa taaaataaaa  14580
taaaataaaa taaaaaaata tggaggccag caggcacggt ggctcacgca tgtaatccca  14640
gcactttggg aggccgaggg gggcggatca caaggtcagg agatcgagac catcctggct  14700
aacacagtga aaccgcgtct ctactaaaaa tacacaaaat tagccaggca tggtggcagg  14760
cacctgtagt ccctgctact caggaggctg aggcaggaga atggcgtgaa cccgggaggc  14820
ggagcttgca gtgagctgag atcgcgccac tgcagtccag cctgggcgac agagcaagac  14880
tctgtctcaa aaaaaaaaaa aaaaatggag gttgggcgcg gtggctcgcg cctgtaatcc  14940
```

-continued

```
cagcactttg ggaggtcgag gcgggcggat cacctgaggt caggagttcc agaccagcct  15000
ggccaacatg gtgaaacctt gtctctacta aaattacaaa aattagccag gcacgatggc  15060
aggcacctgt aatcccagct acttaggaga ctaaggcagg agaatagctt gaacctggga  15120
gatggaggtt gcagtgtgct gagatcgcgc cactgccctc cagtagagtg agattccgtc  15180
tcaaaaaaaa aaaaaaagaa gaaatggaga tacaaactta ctacctacct ccttacaacc  15240
taccctcaca gtattactgt gaataaaagt gtgtgtagca ctgggaacac tattcacaga  15300
gcactcatga atgtttgttc tttgttatta gttactagag aggcaaatgt ctgccagggc  15360
tgaataatat gtgtgaattg gtgattgtcg cacatatcta aagaagtagt tatttttttc  15420
aattaaaact tagtttaaaa accaatataa ggccgagcgc agtggctcac acctgtaatc  15480
ccagcacttt gggaggccga ggtgggcaga tcatttgagg tcaggagttc gagactagcc  15540
tggccaacat ggtgaaaccc tgtctctgct aaaaaaaaaa aaaaagtaca aaaattagcc  15600
aggcatgatg gcaggtccct gtaatcccag ctacttggga ggccgaggca ggagaattgc  15660
ttgaacccag gaggtggagg ttgtagtgag ccgagtttgt gccactgcac ttcagcctgg  15720
gtgacagagg gagacactgt ctcaaaaaaa aaaaaaaaaa accaaaacca atataataaa  15780
taagtggcca gcaatgaaac agaaagtgaa aagttagtga agcaaaacta gtactgtatt  15840
cagataaaga tgctgaatct agatttggtc accagaatag ggtcctttgt ggcaacctgg  15900
gctagtttgg ctgactcacc actgccagga tgaaatttct ttcagtggct actcatttcc  15960
ctttatttta agtccatgct cacagagcaa ccttctgatg cctaattcag cttcctggga  16020
tacttaataa caggaagggt ctggaagtag tacctgtata ggggatatga gtgttctgat  16080
tttaatagtc aattcataag tgtacagagg gtttgataaa tggttaggtc agaaccatca  16140
cagaatgtct acacctcttt ggacattagg aaggtcaaaa acctgaaagg ccaaaagcta  16200
ggcctagatt agggtcattc accaagaaaa catcagcctt gaagagttct ctgggtggtc  16260
caccagtcaa ccttcctttg atcacacctc cttcctcgtt gcttctttaa gcattgacct  16320
gtaatgggta tggaattttt tgctcaccta actccttcct tttacagagg aagaagttga  16380
agcccagaga gatttaatgg cttgcctaag atcacacgca gattttctgt taaccagggt  16440
gatttttcag gtgttccctg ccagacgagg gctttttttcc ttgaattgcc tagagatttc  16500
ttgagatatc cgaagcattt ttcccagtgc agcctggaga aggatgtccc tgtcaacaca  16560
gcatttgtta ctcaatgtta gacattcaat tttctaatta gtatcatgga gcaacagtgg  16620
atgattatct ataaggggtt gcaattccat gcttatgtgc ttacagccca tatagacaaa  16680
tatcagctgt taaaatgaca aggcagtaga gatgtggccc caggacaaag gcatactctg  16740
ctgttagtga acactagttg gccagcaaat ttcacatggg catatacacg gccaactgta  16800
gactttaggc atttataccc attcagagag ccaaactggc aactaaagat cagcattctc  16860
tttggcattt cagctttgcg ttctgttaaa aatcactgct tgcttaaata cctctgatag  16920
ctcttcactg cctgtaggca actctttagc ctagcagact tggtctttag tgctctgccc  16980
ctactctctt ccaccattct ggcctcctgt ctaattgctg cccatatgtg ccatgcacta  17040
gagcttacag acctgctcag cgttatatga gcataccata ctctttatgc ctcagtgcat  17100
ttgcacatgt tgttccttca ggccagaatg cctgttactg cctggcaatc agcctattag  17160
agtctgccaa taccatccca tcttctgtgg aggagccccc cgccaaatcc acccatacct  17220
ctccccacca atcagagact tcttctctct ttgttattct cttcgttatt ctcttcatac  17280
ctcagttata tccatttcag tatttgttta cacatctagc atcactctta gagtgtgaaa  17340
```

-continued

```
ttctccaagt gtggagccgt atctagtttg tctttgtatc ccagagctta gcaaagtgcc  17400
tagaatgtag tgggtgctca gagtgtttgc tgggtgaatg atgtatttgt tgaacgactc  17460
tttggacact tgaataaagt ccatccagta tgcaccatta ccatctcttc gctctacaat  17520
attcttttag gcaagagctt atcttttgag gtgataagat aagctcaaac ttatgtagac  17580
taagacctca gtctgtaaat gtcatcccta agtcttaaac catcaaaacc agggcctcaa  17640
ggaatggcat gccttctgca actgtagcaa cctgctgtgc ttattttgcc gtgtttttca  17700
tttttccccc aaaagctaga gtcccttctc ccatgggcag tgctggaagt gtgctaacaa  17760
attctttctc catactgctt acgattacaa aaaaaaccct cagcatctca tgccagactt  17820
gagttaaggt tgttttcttt tgtgtgtcag ctgtattctg gtcatgactt cctgatgatg  17880
ccctatagag attttgctga gatcagaggg tgctccactg ccatcagtag cactgactct  17940
tgcagaagca ccgtttctga agttggctaa tgtcatccct cacgtttgtt tgtttgaaat  18000
ttgttttagt tccagagata gcactttcat ggaatgacgc tatcttctag aatcactttt  18060
tttttttttt tgagttggag tctcgctgtg tcgccaggct ggagtgcagt ggcacaatct  18120
cagctcactg caatctccac cttccgggtt caagtgattc ccctgcctca gcctcccgag  18180
gagctgttac tacaggcgca cacccccact cctggctaat tttatgtgtt ttagtagaga  18240
cggggtttca ccgtgttggc caggatggtc tcgatctcct gactttgtga tctgcctgct  18300
tcagcctccc aaagtgctgg gattacaggt gtgagtcacc gcgcctggcc tagaatcacc  18360
tttttatacc ataacgtgag caccactgcc gcgtcaccaa ggaaagagag aggcagctac  18420
tgtggggtta caaatgggta agagtggcac caggaaggtg aaagtctcta cttagccaag  18480
gcttaacaaa atgtcaatca ccaaacattt atttattaag ctacgttcag gataagaaga  18540
tgaacaagct atctgtacat tcattttctc gtttgtaaca aggtaatgat agtgatctat  18600
cctgcctgcc tctgagggtt attgtgagaa taaaatgaaa tcaagtggaa aagcacttag  18660
gaaaaagaaa agcattggtt ttcaattgtt agtgtggatc agaaacactg gggcttgttt  18720
aaaatgcaga ttcttagccc cagtctcagc gattctgatt ctgtatatct gaagtgggac  18780
tcaggaatct tgattttcaa caagctgacc agagggtcca atgctgctat tcctttagtt  18840
acactttcag aaatattact gtaaatcaaa tggcaagaat aaaatagtta tttgaggcag  18900
ttttagtatg ttggacctgg agtccaaaga cttgggtcaa actccagctt tgtcagttcc  18960
tagacctgtg accttaaaca gcaaccttct ctgtgaacct tagttccctc aggaacggct  19020
ctggtcacct cctgctgtac tccattgatg actcaccaca taaggctccc tgggagtccc  19080
ccaaaccttt gctctcttaa ctccttttac agcctcctac atctcctgca ggtgctgtct  19140
tctcctcctt tttccaggcc ctgctctgac acagcattca ttctcctctg ggaagggttc  19200
cttcaatgtg tctccaagca catcacaccc aggaaggacc ctgtggccat atctgtctat  19260
caccagatca aactacgtga aggcaggcac taggtactgt cagtgcccag cataggcctg  19320
gcccatacca ggtgtccaca gatgcctagt aaagaaacct atgattcagg acccccatga  19380
tgagcaacta tagcactaga acagtgataa taactaatgt ttataatgca tcttcagttt  19440
acagagggct tttgtactca tcatctagtt tagttcctgc aacaacctct tgaggaatat  19500
agcacaagca ggacaaggga agcccagaga tgttaaataa tttatccaag tttatgctgc  19560
tgggaagggc agcactgaaa ttaaaagaaa agttttctga gctcaaatcc catgcccttt  19620
cctcaatgtg agctctagca aggtattcag gaatcctgcc tctacagttc agagcctcaa  19680
```

-continued

```
attgctgggt atgttgagtt cttgtatctg atttttctag atttcctgcc cacattctta  19740

ctgtctggat atcaggaaag agtttatcaa atgcctgtgg aaatccaaga taaggtctca  19800

tgatgagtaa cccagtgaaa acatgaagtc aagtctaact agtcactact atttcactac  19860

tgctgactcc tgatgatcag ctccttttct aagtgcttac tgtccactta ttccatcatc  19920

tgcctagaat ttatgtgaag gaatcaaagc aaaaggatca taaggcttcc tttttccagt  19980

atgtttttcc tcctttttga aaactgggcc agttagctat ctccattttt atttcatgaa  20040

tacatcccca gcgcctggta tatagtagat atggaacatt acactttgga gatattgcac  20100

ccattctcca gtttctccaa agttactaac aatggttcca tcactgtgcc aacatatttt  20160

ctttttttcaa tatattggga aataattctc ccagtctgaa aatctgaaca catttcatgt  20220

gacttggtat cctcatatgt cttgggcttc caattctcca ttcctagttt caagttcatg  20280

aactgtaaaa caaaggatta gactaaatct ctaaagttct atccagatgc caaattcttt  20340

tctctttcca tgatacctaa gatagatgcc aaatattgtc ttttacctgg tgtttgtgaa  20400

catgacatca cattacagga gtagcagata ctaaactctc actctgtaaa acactgactg  20460

agttccatga gccagatact gaagtgagct tgttcacata tgttctcatt taatgctcat  20520

aaccctgtga agctgggaat tgctgggaca ttttatttat ttatttattg agacggagtc  20580

tggctctgtc acctaggctg gtgtgcaatg gcatgatctt ggctcaccgc aacctccgcc  20640

tcccgggttc aagcgattct cttgcctcag cctccgcagt agctgggatt acggggcaca  20700

caccaccaca tccagctaat tttgtatttt tagcagagat ggagtttctc catgttggcc  20760

aggttggtca cgaacacttg acctcaagtg atctgcctgc ctcagcctcc caaagtgctg  20820

ggattacagg catgagccac catgcctgcc cgggaccctt gttttagaag gatgactgct  20880

gctataatgt agaaagtgat ttggaagagg ggaggagtgg ggcacgaaag atggttagta  20940

gatgggggtg gtaatgctta cctttcagta tttggaggct tcggagtcct caaaaattct  21000

cttccttgat tggagtcctc ccagccaata gagggcttca cacaaacagt ttcttgggtt  21060

ttgaattgtt tgaccagagc tttcttccga caaaaggttg gggtgattca ttcacttacc  21120

acaccttgcc tgaacattca cttggggctg ccggttatga aggctattgt tctccagcct  21180

gtcacagacg ctttgaagac ctgtgcctca gctggttcta aggagtcagt ttgttcagct  21240

ccgtgccagg tttccaactt atgaaatgtg ctggagatta acacctctcc tgccatttta  21300

tccctactat aattgccagt caaaggattc ctgcagttgc ctctggcagc cataactgat  21360

gaatgttctg ccagctgctc tgaggaccta gaagagcagt tttctatcca ggaccagttt  21420

ccaagggtgg gagggtgaaa tatatcctcc agtgtgacat ttcatctccc agtgatgggt  21480

ggcttgggcc ctttgaagtt ggctctgagg aaccacacac ttgggtctga gcagccagca  21540

gcttatcaca tctggtgatc aatccttcaa aggttcctcc tgaagtctga atttttggag  21600

gtcaaatgga ttccacctgg gaggggcttc tgcttcaact caggacatgg ggagaaggct  21660

gttcctcttc caggggggagg cagttttcat ggcattgaga tgtcctctca cttattcccc  21720

acccacccac caagtccttt gtaagaggag taggggggaga ggagagcgcc tgcagcctcc  21780

tgctcacatt cctagacacc gactcactga gcccgtcgcc gctggaacag cagagctgtg  21840

tgaaatgtca agaggagtta tgctcatagg ctccctggcc tcagtctctt tgtggcttgc  21900

atattcttcc attagtactg tgttcatcac atggaaatca gagggtacaa ttaaaagata  21960

atttgctagt cccagactta atttggggcc cccttcttgc ctgattgaat tacaggggaa  22020

cataatagat ttttggtgag aaatagttgt ctgtgtggct gggagaaaga ttgctcccag  22080
```

-continued

```
ctctccagct gggcagccct ttcagtatcc cgtatgttat ttccccactt ccagcccacc  22140

tcacctcctc tgtggccctt gtgtgtcccc tcggctagga tcctgacctc ctgctcaaga  22200

gtttaaactc aacttgagac ccaaggaaaa tagagagccc tctgcaacct cataggggtg  22260

aaaaatgttg atgctgggag ctatttagag acctaaccaa ggcccagaca gagagagtga  22320

cttgctaaag gccacatagc tagcccacag tagttgtaac aatagtctta atgatattaa  22380

tggctaacat ttatcaacct ttaatgtgtc ccagactttg tgccaagggc ttacatgcag  22440

tgcattgtcg cattcaaacc cagacagtct ggctctgggc ccaggctgag ctttggtata  22500

gcatggtaga acgttgtcta taatgtctag tctgggttca aatcctggct tcacttctca  22560

catttacagc tgagtgacct caggcaagtg atttaacctc cctgtacctc agttgcttta  22620

tctgtaaaga gaaaaatcac agcactgtgg aatagtgggg gttaaaattc attcatacaa  22680

gtagtgctgc aagcaatgtt taatacaggg tgagcacctg ttcagtgctt ccttcttctg  22740

gctgcctctg ggctagagt gtggtgtctt cgtggtatag atagatagat atggctgagc  22800

tctgcacaaa caccaagagc tgttcttcac tattagaggt agtaaacaga gtggttgagc  22860

tctgtggttc tagaacagag gccggcaagc tatggcccat tgcctatttt aatacggcct  22920

gtgattgatt gatttttttt ttctttttga gacagagttt cactcttgtt gcccaggctg  22980

gaatgcaatg gcacgaactc agctcaccgc aacctctgcc tcctgggttc aagcgattct  23040

cctgtctcag cctctcgagt agctgggatt acaggcatgt gccaccacgc ctggctaatt  23100

tttgtatttt tagtagagac agggtttctc catgttggtc aggctagtct cgaacttcca  23160

acctcaggtg atctgcccgc ctcagccttc caaagtgctg ggattacagg cgtgagccac  23220

catgactggc ctgattgact gattttttta gtagagatag ggtcttggtt tgttacccag  23280

gctggtctca aacttctggc ttcaagcagt cctccctcct tggcctctcg aatgctggga  23340

ttataggcat gagccactat gcctggccta tatgacctgt gatttttaat ggttagggga  23400

aaaaaagcaa aagaatgctt tgtgacatgt ggaaattaca tgaaactcaa atatcagtgt  23460

cccagcctgg gcaacaaagt gagaccctgt ctctacaaaa aataaaaaaa aataagccag  23520

ggccgggcgc agtggctcac acctataatc tcagcacttt gggaggccga ggcaagtgga  23580

tcacctgagg tcaggagttc aagaccagcc tgaccaatat ggtgaaaccc tgtctgtact  23640

aaaaacacaa aaattagccg agcatggtgg catgcgcctg tagtcccagc tacttgggag  23700

gctgagacaa gagaattgct tgaacctggg aggcggaggt tgcagtgagc caagatcgcg  23760

acactacact gcagcctggg caacagagcg agactccgac acacgcacgc acgcacacac  23820

acacacacac acacacacac acgctgggta tggtggccag cacgtgtggt cccaggatgc  23880

actggaggct taggtaggag gatcacttga gcttaggtgg ttgagactac aatgaaccat  23940

gtttatacca ctgcacttta gccagggcaa cagtgtgaga ctgaatctca aaagaaaaaa  24000

aaaaaaaaga aaaaaatctt tccataagta aatatctgtt ggaacatagc catgtccctt  24060

agtttatgtt ttatatatgg ctgcttttgc cctataatga cacaattgag tggccacgac  24120

agtctgtatg gcctgcagag cctaagatat ttgctctctg gccctttaca gaaaaagtgc  24180

cttgacctgt gctctagagc catatgtacc aggtttgaaa ctcagcctca cagctgggtg  24240

tgatggcacg catctgtagt cccagctact ctggaggctg aggtgagagg atcacttgag  24300

tccagaaggt cgaggtcaag attgtagtga gccatgatgg catcaccgca ctccagcctg  24360

agtgacagag agagaccctg actcaaaaaa aaaaaaacaa aaaaaaaaaa caccctcacc  24420
```

-continued

```
acttatcagc tatttgtctt gagaatagtg acataacccc tcagaaccta tttcctaatc  24480
tgttaaatga ggctgatgac gtttcctcct tttactggca atttaaacat gatggataat  24540
aaatgctaag cacttaacac agggcctaga agatattaac tgctcaataa atggtagctt  24600
cttaacagta ttcaaaccca tgtgctctta tcacatgcat tgttgtccct gtgtccagtt  24660
ggtggaatgg gaaaaggctc ccttgtaacc ccatctacca tctttatcag actttcctgc  24720
catggttcac agtaagagat agaagctgca cggtgacttc tggctcttta caatggtgag  24780
cggtgtgtgc ctggtaaggg agagctgatg tcactgcccc aaatccagta gtgagatctg  24840
agtgttctgg tttcctccag cagccttgct ttttccttta caatcctgca ggcagggaga  24900
caagggcttt ctacatggta ggctctggtt tggtcatcgt cacaactggg ggctgttcag  24960
gtgggctccc attccagata cctaggctta tcaatccctt ttggcacccc aggccttttt  25020
ctccctcatg ccccattttt cagtttgaaa agcatggtta tcacaggaca agtagaagaa  25080
gctccactgt ccactgaggc caatggatgg tgttctgcat gtgaacactc agtgaatagt  25140
gagtgaatga gagtaacctg ggctccatcc tatttgcaga gagctttgga aaagattttt  25200
ctccttaaag agccagaatg aagcctggta gtgggagagc tccagctcta gagtcacatg  25260
agcctacatt taaattccag ccctgccact gactcccttt ttgaccttga gtgagttacc  25320
taatctctct gtacctcact tttcttgtct gtagagtggg aataattcct gtctcagaga  25380
aataaaagag tgcatatagt gtttgccaca tggagacaca tcaggtgtag gttaatactc  25440
tgggccttgt ttccttattt gcaacacagc cctgccctgg agtggaagtg gcacctccca  25500
ttggtcagct cttgaggctg tccccaggac aggcagaggg agggaatgaa tgggagccct  25560
agtgccagga cagaacagat ggcagctcag agctaggatg gctctctgga cctgtctctc  25620
ctaccagagg tccccccgtc tggtgtggct cttcctggac ctggcatcct ctgctttttt  25680
tttttttcca cctccaagca gaattactgt cctgtaggca gctcctctgc ttgaggacat  25740
ctggggccag atatgttcac actctatcct gccttgccct tccctgagct caggatggac  25800
gctcaattgg tcccagttat tgtctgcagc gcctgcctgc agcctcgatc cagcccagct  25860
ccaccccttg cctgcaaggt ctgtttccta acagctgctc caaccacaca cctcggttct  25920
gcgggagccc ctcctcttcc tccctccctc cctcattcag ggtgggact gaagaagaag  25980
gctaacttga cagcagcgct tctttcttag ctagtcaccg gcccctgctc aagaatgcca  26040
gtgtgtgtgt agcctccaca gagaggtcgt tttctcggag tccagagggg ccgcctgagc  26100
ttctgagaac tagggaggag ccatcccagc catgagcccc tgtgggaatc tgctgggggc  26160
caagtggcct ggagtcctca ggctcccgca gctgctccgg agggagaggt gagctcaggg  26220
cagcctgcct gcagccagag gtgccgggag ccccgggcct gtcatggtgg ccatctacag  26280
ccggcctgag gcagtcacag acggatttgc agctgagcct gtctatctgg tgtgggaaga  26340
agatggggag ttacttgtca gtcccggctt acttcacctc cagagacctg tttcggtgag  26400
ttggtctccg agttcccctc tccatctctc ctggcccctg gtcctgagag gagggtggtc  26460
tccctaaatc tccttctcac ttagtccttt accatcggtt ctgccgggca gaagccagcg  26520
gaggttatac ccaaggagaa tcggccttgt gaggtacccc cattatgtcc tggaagtggt  26580
gaggggaggg atatacccag aaggaacttc ttagggagct ccagctcccc ttctatccca  26640
gacaaacctg aaggagcctc caaaagatgc cactgacctg cccattgtag atgttactgc  26700
ttccgggggg aatagcccaa atagagtgct gtttccagct ctcacatgtc ttacctgcgg  26760
gccatgctgc ctgcccagga atttgtccca acaagcagga tgggcaggtt ttgccaaact  26820
```

-continued

```
gtggaaactg gcaagtcctg ggtgtgggta gcctggtaca cagtaggcac cttataaacg  26880
tttgttctct taatggcagg cacatttgcc tctggccttg aagggcttct gagctcccag  26940
gtgaatgtag ttgctgggga aagacctggg cgagtgcttc taagactgga gcaatgggct  27000
ttagagtgtt cctgagctgc tgggccagcc cccacacctc ctcagtccct aggcctaagt  27060
acctccacga gcctctctct gtggggcttc tcagagggag atgtggaaac tctacctcta  27120
acctggcttt ctttgctcat tgccccactc cacctcccat agaaactccc caggggggttt  27180
ctggccctct gggtcccttc tgaatggagc cattccaggc tagggtgggg tttgttttca  27240
ttctttggga gcagcctgtt gttccaaaaa ggctgcctcc ccctcaccag tggtcctggt  27300
cgacttttcc cttctggctt ctctaagcta ggtccagtgc ccagatcttg ctgccgggat  27360
actagtcagg tggccaggcc ctgggcagaa aagcagtgta ccatgtggtt ttgtggaatg  27420
accggaccct ggtagattgc tgggaagtgt ctggacaggg ggaaggggga agggaactgg  27480
tcctcaatgc tgactctacc aagcgccctg ctagacactt tatcctttaa tctctcaaca  27540
gcctaaagag attatatatc cccattttac agatgaggca accagtttca acagagttaa  27600
catatggagc ctcactgggc agctttttct gtcttcctga ctttctctca tccttcaggg  27660
ggctgcaggt ttgttttctt ctcctagtgg agaggaaatt ctcaggtttg ttttcctctc  27720
ctagcagaga gtaaaaaaag ggatagtttg cctgacttgt tgaaggtgtg gctgagattg  27780
ttttctaaag agccaatgga aattgatctt gagtttagga gaaagctttt acatgtggaa  27840
ttaagatgcc aagtgttgaa gtagccacat ttcaggtcct cattaatttc tcttaatcct  27900
gggaaggcag cttaggagaa gggttgttcc tttaggagcc aggaactata cccctttttac  27960
ccttggagag gcagggaagc cagggaggac acaacttctc aggaagagga gaagctagag  28020
cagatagtga actctcaacc tgaaccttta agggccagac cactaatgcc acccaagtcc  28080
acctgccgtt tgtcttgttc tgtcccaggc tttctggaga acctgatctt cttgccccta  28140
cccccaagct ccgtttgccc agctagagtc tggggggtac tgactgactt tcgtagacat  28200
tcttcccttc cccaaataag aggccacatt cctgaagtca cttctgaaga gatagctgcc  28260
acacagggct ctttcccccc agggagggac cacccagacc ctctgctctc ccaggtatcc  28320
gttaccacat cactacctgg tcagaaagct gtttctgcca ttagcccctc cctcttttat  28380
tataggatat cctcaagggc tcctctttgg gcctcagttt catccttggc agaaagtaga  28440
agctagactt cttgggctcc tgaacagggt ccttgctgga ttctgtgaaa caaattaagt  28500
tcttgaccct aggcctctgg gggagtacaa agtctatggg agttctgggg ctgtggttgc  28560
aaggaaagtg acgcaaccag attccatggg gacatgatca ggcgtgacat gtgagggagg  28620
aagagggagc aagggaatga agaatacaac ttctgtgtcc catacacccc tgcctgacag  28680
gccatacata ctcagcagag aatgcactgt ctttcctacc acactagcgt gaggagtgag  28740
ctgcaattac cactgtgctt ccaagtaaga aaatacctca aattggaatt tacaaaagag  28800
gtaaattagg gagtggcttt tgtcggacat ctttaaagca tttttctttt tatagaattt  28860
cacttaatgt ccaatactga tttaatgagc ttgggtttac acattatctc ttgaagaaaa  28920
caaatgaacc tttgtgttcc aaagcaatcc atgtttaaag ggaaaaaatt atgcataact  28980
ctgcccagct tcacagtaac ctttggcagg tgccttaggt cctctgggac tcttttcctt  29040
atctgaaaaa tgaaggactt ggatcaggtg aatggttccc agctctgcaa cttatgtggc  29100
tcctcagagg cacacaagct cttttccatt atttgccaaa taatggaggc cctgtcttta  29160
```

-continued

```
actgcagtac aactacacaa aatacttgaa actacagtct tcctggtttt tggttggaac  29220
tgaatcagtg cactctagca acacttattt cttgctgttc gtaggcttca ttatgtgttt  29280
ggttaatttt ttaaaacaac aataacatat tccataataa ttacagctta attggcagac  29340
tgtttcagtc tataggatct gcaggaagga ggagtaataa agggattttt gactgagctc  29400
ttatggaaca gagtctctct aggcccctgt catatctgcc cttctgggcc ctggggaaaa  29460
gttggcatcc ccagttgtgg tgctctccag gtgccctcag gctgtggtgg agggagcttc  29520
ccattctctc cttcagccca ctcaattcag aggctagggg ctgaaagaag cttctctaca  29580
actggctgtt cactgggagg ttaagggatg accatccagc caggccttcc tcaggacatg  29640
ggagggctta tgctttaaca tgtgtaaatc cactgcaata atgactggtt cttttacccc  29700
ataaggttga gaatttacct gtaaacattt ttgtctgaag aatttggatg taagtgaggg  29760
ctgggcctct atcttatctc acttggcttc tctcagcaca gcaccttgcc tgcttgttct  29820
tacacatcct agatgcacag taactatttc ctaattatta gaaatctatt agaatcaatt  29880
gatttcagct gggcttggtg gctccttcct gtaatcccag cactttggga ggctaaggct  29940
ggaggatcac ctgagtccag gagtttaaga ccagcctggg caacataggg agaccctgtc  30000
tctacaaaaa ataaaaaatt agccaggcat ggtggtgtgc acctgtagtc ccagctactc  30060
aggaggctga ggcaggagga tctcttgagc ctgggaggtc agactacagt gagcaatgat  30120
tgtgccactg cactccagcc tgggtgacag agtaagactc tgtctcttaa aaaaaaaaaa  30180
aaaaaagttg atttctattt ggatagataa ataattcatt ttaggacctt tctttttcac  30240
ttacagaaat ctgtttcatt ctgggctgag aagcaggtcc atattgctag gcataggaga  30300
aaaaggggtc tgtctgcatt tgcccttggt ggtctcaaat tggggaggga aagaaatgaa  30360
cacttactgg ctaccttctg tgagccaggc atcatgcaag acatctgtac ataatttaat  30420
tctcataacc ccataagata ttattagcaa tgtacaagtg aggaaactga ggctcagagt  30480
catgaagtaa ctggccttgg gtgacacaga tggtaaatgg cagagaagga atatggatcc  30540
aggtcttgaa agagaaaatc tcaactgatt atctttttta aaaaactcat atgttctctg  30600
ctgactcaaa aggtctctgt gtggatctgg gttgacccac tgaactgacc atcagggttc  30660
catgcacttt gtatctgccc aagccctcag aacccctcag taatgttttg gaagatgagt  30720
tttggaggtt gtccttaggc atagcctcag cgtatgtagg cctctaggtg atctccccta  30780
acctgaggat ttcagctcaa ttcactctgg ctcctcagga cagtgggatg actggttcag  30840
acctcagctt taccacctcc cagctgggta ctcttctacc tacagccagg gcagattttg  30900
actttcactt gaaacttcca aaaattgaaa ggtagaaaaa cagccttggc tttgggaaga  30960
acgtatgatg tccatggcct ctaagcatct gaggtgggac atgttcgagt agcaccttac  31020
agttccaaag tgtgttctgg gttctttgtt taaaagaaca gagactgctg gggaattgaa  31080
cactgtgaag tatatgaagg aggagaattg tgctatttaa cattcagtac ttgggctaaa  31140
ggagaagcat cacgaagtgt taacactcaa agggtcttga gctgtcaggg ctccagcttc  31200
cttattttca caggtgagaa tcctgaggct cagctgttga gatgtgctgt ctcactccgg  31260
tgacatagta cagtggatgt ggctttgcag ccaagcacac atagcttcac attccagctc  31320
catcaattat gtattgggca gctttgcaga atgatttgac tttaactctg cttttcagtc  31380
ttctgtaaaa cagggataat cctgctaccg tagggttgtc aggattagag ataatataaa  31440
taaggtacct catataggac ctggattatg gctggcattc aataaatagt agctgttaat  31500
tgatagctaa gctagaactc tgaagtctac catggcaact tcttaagtgg tctgagaacc  31560
```

-continued

```
cagttgtgtt ctgtggcaaa acacagctta gggatccata cccagccctc ctgtcagctg  31620
ttcaccttcc agttcttcag agacatgtgt ggcagtgact ttggccacat agctggctgt  31680
gccctttaaa ggcattcctt gacacagata tgtggactgg tgacgttgct ctccagccag  31740
gtgttcttcc cagcaggctg gcctggctgt ctcctgcatg cctgtacttg tttgtctccc  31800
tgctccctct cctgggcctg gccagagcta cttgcagcaa acaaaagcag gatattggca  31860
atggaaagga gggtgtgttc tggtgctccc atgccctgcg gcgcacatac cattgcaagg  31920
gcgtaacaga gcccaggcct gcatttgggt gcaaataagt ctgcacacag aagaaaagaa  31980
ggacctggtg accaggagcc atggaaccct tgtgctcccc tacctgggct actggttctt  32040
gccactccta ccattttcag tttggaaata tttgttaagg ctttgctctt ccaggtcctt  32100
tgcttggtgc tgagtctacc aagagtaagt gggatgctgt tttgtcctc agggagctaa  32160
cagtctagtg aagaagaaag atggttgccc aggaacttct aagtcagaag gcaggaggca  32220
agaaggaagc ccctgctcct actgccagcc ctctgttggg caccccatag ttcttcagaa  32280
ccacatttaa tcctcactgc aggccaggca tagtggctca cacctgtaat cgcagcactt  32340
cgggaggcca aggcgggcag atcacttgag gtcgggagtt cgagaccagc ctcaccaaca  32400
tggggaaacc ccgtctctac taaaaataga aaaattagcc gggtgtggtg gcatgcgcca  32460
gtaatcccag ctactcagga ggctgaggtg ggaaaatcac ttgaactcgg gaagcagagg  32520
ttgcagtgag ccgagattgt gccactgcac tccagcctgg gcgataagag caaaattcca  32580
tctcaaaaaa aaaaagaaaa aagaaaaaat cctcactgct accttgaaag taggtgatga  32640
cattgccatt tcacaaatga gaagtgaagg ggctagccca agatcactta ggtggtaaat  32700
ggtggtgcta agattagaac ctcagatcat ctagggaaaa acacagatat gcacagagtt  32760
aaggggaccc agggtattgt ttgtcctctt gtttcacagg tggggaaaca acccagagag  32820
ggaaaggggc ttgtccaagg caatttagca cccaagaact tgaacccata tctctctcct  32880
cctcatttag agctcatccc acatgtatct tatattgaga ggagtgtgag ccacatacca  32940
agaacagtct tcccctctgc ctccaacctc actgtgcagt tttgagacac ttcacagcca  33000
tactcttcat gccataccca gcccttaaga ccctgaagtt ccccttccat aagacaagta  33060
ggaaaagcta tagggtaaaa atagccatca gtgtttgttg agcacccagg aggaattggg  33120
cactccagaa agataaaggg attctcaggg acttgcttct ctagacttcc ctagctcagc  33180
tgcttcaact cattcctgcc cctcttctct acctcccgca gtgctcagaa gtagtagaac  33240
tcactgtggc ctctcacctt gcattgttga gttttattta gactttctct tcctcaactc  33300
ttcataagct catgaaaggt gaagtagggt gccctgtgta tttatctttt atatctgcag  33360
tgcttagcaa gttataataa tgcacttgcc tggcaaaagg ctttctctca tacattagct  33420
tatttcctct tcacattggc tctttgtagt aataggatgc tattagttat tttcaatgag  33480
agaaagctac taagagaagt tgtccagcta gtgacagtaa gtggctgata aagtgagctg  33540
ccattacatt gtcatcatct ttaatagaag ttaacacata ctgagtttct actatattgg  33600
gtcttttttt tttttttttt ttttttttta gagacggaat cttgctctgt tgtccaggct  33660
ggaacgcagt ggtgcaattt tgggtcacca caacctccgc ttcccaggtt caagcgattc  33720
tcctgcctca gcctcctgag tagctgggac taccagtgca cgccaccacg cccggctaat  33780
ttttgtattt ttagtagaga cagggtttca ccatgttggc caggctggtc ttgaactcct  33840
gaccttgtga tctgcccgcc tcagcctccc aaagtgctgg gattacaggt gtgagccacc  33900
```

-continued

```
gcgccctgcc tatattagga cttttatata agctatctct agctagctag ctagctagct  33960
ataatgtttt ttgagacaga gtctgactct gtcacccagg ctggagtgca gtggcgtgat  34020
ctcgactcac tgcaacctcc acctcctggg ttccagtgat tctcctgcct cagcctcccg  34080
agtagctggg attataggtg catgccacca cgcccagcta atttttgta tttttagtag  34140
accaggtttc accatgttgg ccaggctggt ctcgaactcc tgacttcaag tgatccaccc  34200
gcctcggcct cccaaagtgc tgggattata agcataagcc actgtgccca gctgctctct  34260
atatttttaa tacatattat ttccattaat tttcacagca gttcatttta tagatgagga  34320
aactaggcca gagaagtaaa atatcttgcc caagatgatg taactagtaa gtggcaggat  34380
caagattcaa accaagcaat gttcaaacct cttggaagca agaatgtggc cactgtggaa  34440
ggtgcaaggc cttgacaaca agaataggga aaagaaggaa ctagaaggaa agagatggca  34500
tgggctcagc aggccaggga gctcttagct gtgtgtgttg ggaagctcag aagggaggaa  34560
gaggttgtct gtgcaggtaa gtcctgagaa cacaccagac ttttgagagg tggagcttca  34620
tagccaggtc attaggggag aagggagcta tagattttt tttttttttt ttttttttt  34680
tttttttag agacggggtc ttactatgtt gcccaggctg gtcttgaact cctgggctca  34740
agtgatcctc ccacctcagc ctcccaaagt gctgggatta gaggcatcag ccaccccgcc  34800
cagcgagcta tggatctaac atgtacatct tacacagtgc taatagaatg ttgggtttct  34860
tccccaatat tttattttga aaaaaaattc aaatatatag aaaagttgaa aaatgtagtt  34920
caaagaacac ctacatacct ttcacataga ttcatgattt gttaatgtta tgccactttg  34980
tatatatctc tctccctcct atctgtatac ttttatttat ttattttgc tgaactattt  35040
cagagtaact taaaggcatc ttgattttac ccttgaacag ttcaatatgt ttctgctaag  35100
aattctccta tataagtcag atatcattac atctaagaaa attcacggca attttacaat  35160
ataatattat agtccaaatc catatttcct cagttgttcc aaaaaatgtt catggctgtt  35220
tccttttta atctaaattt gaatccaagt ttgaggcatt gtatttggtt gctgtgtctc  35280
tagggttttt aaaatctgtg ccttttcttc tccccatgac tttttagaag agtcaagacc  35340
ggttattctt atagaataac ccacattcta gatttgcctg attagttttt ttatacttaa  35400
cgtattttg gcaagaacat tacattggta acgctgttgg tgatgggtca gttttgaaga  35460
gtggagatga ttaaactgct tttgttcatt gaagtatctg tcaagaccag agatccttaa  35520
ctggtgccat aaataggttt cagagaatcc tttatatata caccctgtcc cccacctaaa  35580
ttatatacac atcttcttta tatattcatt tttctagggg aggcttcttg gcttttatca  35640
aattctcaga gggccccaag acccaaagag gttatgaaac actagtctgt ccactgaggc  35700
aggcaacaca gagctggttt ctggggcctt gttcagtctg aaccagcttc ccttggggag  35760
atagcacaag gctgtaactt tgccccatct tggctttgga tcaaagagga ctgtccattt  35820
tgttgtcata cctaggaacc agggacagct tatgtggcct ggttccaggg atccaggaga  35880
atttcagttc ttgtcttgcc tttcaggtgt tcagaatgcc aggattccct caccaactgg  35940
tactatgaga aggatgggaa gctctactgc cccaaggact actgggggaa gtttggggag  36000
ttctgtcatg ggtgctccct gctgatgaca gggcctttta tggtgagtga atcccttcat  36060
atctgcccct cttggtcttc agagtccatt gacagtgctt ccagttccct gtggcctgtt  36120
aatcttttag tctttccatc agccagggca tctcccttta tttattcatt cattcaacta  36180
gcaggtatca attgagcacc tactaagtga aaggtaagat ccttccctca aagacttaat  36240
agttgaacgt tgggagtggg aggagaggca ggcagagagg agacacaata tagttggata  36300
```

-continued

```
aggacctcca aggagagtgt tacaggctga gaggaggata tacttaggtt gtctttaggg  36360

aatcagaaaa ggagactctg gaataggctg gcagagagag gggctacctc ctatacctgc  36420

tctggacaaa cgactttaag catagtgaca gatttgccaa ccctgtattg gaagaactga  36480

tcttttttag tggggatgat tacttctggg gatttcttct cataactgag accaaaacag  36540

ttttgtgcag tctcagaaat gacaggaggt accaatctga cacttccttt ggaagctcta  36600

gggcagagag tgaaagagtg gattttgacg ggggccttgc ttggaggtca ttcacccacc  36660

cctgtcctca ctccagcaac agtgataact cacttccttc ctccctttgt acacccttct  36720

ccccacctgc tcacaggtgg ctggggagtt caagtaccac ccagagtgct ttgcctgtat  36780

gagctgcaag gtgatcattg aggatgggga tgcatatgca ctggtgcagc atgccaccct  36840

ctactggtaa gatagtggtc ctttgtctat cctctcccat ataagagtgg ctggcgggga  36900

gggacagtgg cagggtgagt tgggcagaag gagtgttagg gtagtcagag cattggattc  36960

ttaccacagc agtgctctta accagctctt taacttgtaa gcagaatgat ttacacatgt  37020

ctctaccctt tttccttacc aaccttgaaa atgtcttcac tctgccctgc aatcctccca  37080

gtgggaggca ctcttcaagg acgatcccag aacattaaag tcaaagaccc cttagagctc  37140

accctgtcca accaccttgg ttgataaaag aagtcagcct ggggcccatg gaatagaata  37200

gtacaagggc aaggttctca ttgtgagtca aaggtagagt gaagagaacc cagaccatct  37260

caccccaacc caggccagtg tttttccaaa tataccactt gctgcagatc tagctcagca  37320

cccccagtcc cagcccaccc tgagaaccca ggctcctcat tctgagcagc cagctagaat  37380

catgacaaag agggtggtag tgagactatg ggtactgttg cttaaagcca catggtgcag  37440

tggttgctgg ggggcttctg tgtgggactc tagcatctta ttcccccctg tgccctctcc  37500

ccagtgggaa gtgccacaat gaggtggtgc tggcacccat gtttgagaga ctctccacag  37560

agtctgttca ggagcagctg ccctactctg tcacgctcat ctccatgccg gccaccactg  37620

aaggcaggcg gggcttctcc gtgtccgtgg agagtgcctg ctccaactac gccaccactg  37680

tgcaagtgaa agagtaagta ttttgagaac ccttcagcag gggttcttga gcagagtctg  37740

taaatgggcc tcagagggct tagacctcca aagtctcatg cagaactccc tttattctca  37800

tctcatatct ttctcctgga ccccactatg ctgtaaccgt acctgggcct tggcacttac  37860

tgttctctct gcccaggcta cttcctaccc gatacttaag gcaagaatca ctcacctttc  37920

aggtgtcagg tttcaggtca tgtttgctct ttgaaatcat ctggcttgat tatgtgtatt  37980

agttgtttat cttctatccc ctccactaga atgtaaattc cagaagaaac ttgctgtctt  38040

attcagtgct gcatgcccag ggcttggaag agtacctggc atatagtagg agttgattga  38100

ttattatttt gtcagtcgag agaatgaatg gagaaaatgt ggtccatggc ccaaaagaag  38160

ttaagaccct atcctagatt caggccagag accagatgga gaaagagtct gtgtctatct  38220

aataccagta atgtcgtacc tctggccgct taccatgtaa atattgattg tgtatctacc  38280

atgtgttgga cactaggcta gtgcttgcac agcaggtgaa agatactaga gtttgggaag  38340

tcaggaggag ctaaggtctg ttctacaacc ttattagatg aagaggagag ggaattgtgt  38400

tcagggcaga gggagaagca tttctccaaa agtaggagtc ttaatcatgt ctgatgtagg  38460

ttgagtgtgg ccagaaaagg ggctgttaag tatagagggc ctggattatg aaaatccagc  38520

agatccattg agagtttaag cagcaaggtg ttgtgaccaa gttaacattt tagaaggatc  38580

actggtatgg aggttggatt ggagagggga aagcctaaag gtatagagac tagttaggaa  38640
```

-continued

```
gctattgtag gctgggcatg gtggttcatg cctgtaatct cagcactttg ggaggctgag  38700
gtgggaggat tgcttgaggc caggagttga agaccaacct ggccaacata gcaagacccc  38760
gtctctgttt ttcttaatta aaagaaaagt ccagacgtag acatagtggc tcacgcctgt  38820
aatgccagca ctttgggagg ccaaggtggg cagattgctt gaggtcaaga gtttgggatt  38880
aggccaggcg cagtggctca cgcctgtaat cccagcactt tgggaggccg aggtgggcgg  38940
atcacaaggt caggagatca agaccatcct ggctaacaca atgaaacccc gtctctacta  39000
aaagtacaaa aattagccgg gcatggtggc ggacgcctgt agtcccagct actcgggagg  39060
ctgaggcagg agaatggcgt gaacctagga ggcggagctt gctgtgagca gagatcacgc  39120
cactgcactc cagcctgagc gacagagcga gactccatct caaaaaaaaa aaagagtttg  39180
ggattagcct ggccaacatg gcaaaacccc atctctacaa aaagtacaaa aaaattagct  39240
gggtatggtg gtgcgcgcct gtaatcccag ttactcagga ggctgaggca tgagaattgc  39300
ttgagcctgg gaggtggagg ttgcagtgag cccagatcat gccactgcac tccagcctgg  39360
atgacagagt aagatgccat ctcaaataaa aattaaaaac aaagtttaaa aaaaaaatag  39420
aagctattac cgtgatccag gtaagagatg tgaataacta caatgatgga aagaaggcag  39480
agttcttaga gatgggagta ggagagatga gggaactcca gattgggaag atgatgttca  39540
agtttctggc ttaggccaca gggtgagtgg caattccctt cactgagatg gggcatcctg  39600
gaaaaggtgt tgcctttctg tgtgggtatc ctgggcccct taggggccac tggtggcctg  39660
ggacctggta aaccttccct gcacaagcag aattggtcaa gcaggttttt aggacatctt  39720
taccctgcct caactcttgt ctggcccagg gtcaaccgga tgcacatcag tcccaacaat  39780
cgaaacgcca tccaccctgg ggaccgcatc ctggagatca atgggacccc cgtccgcaca  39840
cttcgagtgg aggaggtaga gtgtgtgtct aatctgtctt gtgagggtgg gacatggaac  39900
agatcctctg ggaaatcagg ctgtagcctt taccttttcc tacccccagc ccatctcttt  39960
gtcttagcat tgagcctgtg accactggtg acctatttca gcgtaacagg ttcccagggt  40020
agcagggatg gttgatggac gggagagctg acaggatgcc aggcagaggg cactgtgagg  40080
ccactggcag ctaaaggcca ccattagaca agttgagcac tggccacact gtgcctgagt  40140
catctgggtt ggccatgggt ggcctgggat ggggcagcct gtgggagctt tatactgctc  40200
ttggccacag gtggaggatg caattagcca gacgagccag acacttcagc tgttgattga  40260
acatgacccc gtctcccaac gcctggacca gctgcggctg gaggcccggc tcgctcctca  40320
catgcagaat gccggacacc cccacgccct cagcaccctg gacaccaagg agaatctgga  40380
ggggacactg aggagacgtt ccctaaggtg ccacctccca ccctggctct gttctgtcct  40440
atgtctgtct ctcggatgaa gctgagctgg ctttcagaag cctgcagagt taggaaagga  40500
accagctggc cagggacaga ctatgaggat tgtgctgacc cagctgcccc tgtggggatc  40560
acagtttaca gccagagcct gtgcggaccc agctgtctgc caggtttcct tagaaacctg  40620
agagtcagtc tctgtccact gaactcctaa gctggacagg aggcagtgat gctaaaccct  40680
gaagggcaac atggcctatg gagaaagcat ggagctcaga gcctggagta cgggcacaga  40740
taggattgaa taaattgtgt agaaagactt tgaaaacaat aaagcaaaag atgaatgaac  40800
gttttttta gacttgaggg accaacaacc cccaaacccc agattctgcc aggtccatgg  40860
ggaaggagaa gttgccttga gtggaagccc caagtaggga gacttacaga aaagaagtca  40920
agagcactgg ctcccaggca gaaatactga taccctactg ggcttcagg ctgagctcct  40980
cccttcacaa atcacttcat ctctctgagc ctgtttctgc atctgtgaca taagatggta  41040
```

-continued

```
agataaaggt ggctgtctca ccaattatgt aaggattaaa tgtggaaaag gacataaagt  41100
tgtatagtgc tgccataggg acagtgttca gtaaacgtga cacattctta gtatcactaa  41160
gaatcaggtt cttggccagg caccgtggct catgcctgta atcccaacac tctgggaggc  41220
ctaggtcgga ggatggcttg aacacaggag tttgagacca gcctgagcaa catagtgaga  41280
cactgtctct acaaaaaaaa aataataata ataattgttt ttaattagat gggcagggca  41340
ctgtggctca cacctgtaat cccagcactt tgggaggcca aggccggagg attgcttgag  41400
gccaggagtt caggagcagc ctgggccaca ttcctgtctc tacaaagaat aaaaaagtta  41460
actgggcatg gtggcacatg cctgtaatcc cagctactca agaggctgag gaggaggatt  41520
gcctgagccc aggagttcaa gactgcagtg agccttgatc acaccactgt actacagctt  41580
gggcaacaga gtgagacctt gtctccaaaa aaaaaagttt gttttttttt atccactctc  41640
ctcaccaaac aaactgagta agttagagcc ctctcagctg gcatgtgttg gaaacagtgc  41700
cctctcatta aagtgctgcc ctcactccca ttgcctcttg gccttggtca gtatgatgaa  41760
attagtggga ggcagggcaa cagagggcag ggaagagcta gaaatccatg gcctggaaaa  41820
gggaagattt gggagtggcc aggtatctgt agagccacca tgcagaggag gggggcagct  41880
agccttgtgt gctctggtgg gcatggtcag caggaggcag agcaaaagga caagggtaag  41940
taaacctgta ggtcgggaca agccaagagc catccagcgt cagtcctctc tgggtagccc  42000
aagtaaagca ggagcatacc ccagagagaa agttcgcagg gctgttcacc tgcagtgctg  42060
tggacttcaa ccttcttgtt ccttcttcag taagtgaaaa taacagtcat tgaccatgac  42120
tattatcgac cgcttttgaa aatgtaaaca tagtgacttt attgctgtaa aaatcatacg  42180
tgtttatcat cttaaaattc aggaaacatg gacaggtaca aagatgtgca aaatatcatc  42240
caaaatccca tttgctggcc aggcacggtg gctcacgcct gtaatcccag cacattggga  42300
ggccgaggcg ggcaaatcac ttgaggtcag gagtttgaga ccagcctggc caacatggtg  42360
aaaccctatc tctactaaaa atacaataat taggctgggc gcagtggctc acgcctataa  42420
tcccagcact ttgggaggcc gaggtgggcg aatcacaagg tcaggagttt gagactagcc  42480
tggccaatat ggtgaaaccc catctctact aaaaatacaa aaattagggc cgggtgtggt  42540
ggctcacgcc tgtaatccca gcacttaggg aggccgagac agatggatcg cgagatcagg  42600
agttcgagac caacctagcc aacatggtga aaccccatct ctactaaaaa aatacaaaaa  42660
ttattcggtt gtggtggcac acgcctgtaa tcccagctac ttgggaggct gaggcaggag  42720
aatctcttga acctgggagg cagaggttgc agtgagtgga gatcccgccg ttgcactcca  42780
gcctgggcga cagagtgaga ctccatcaaa aaaaaaaaaa aaaaaaaaaa aaattagccg  42840
ggcgtggtgg cgtgcaccta tactcccagc tacttgggag gctgaggcag gagaatcgct  42900
tgaacctgga aggcggaggt cgcagtgagc cgagatcgtg ccattgcact tcagcctggg  42960
cgacagagcg agactctgtc tcaaaaataa taataataac aataactagc cgggcctggt  43020
ggcacatgcc tgtagtccca gttactcagg aggcggaggc atgagactca ggtgaactag  43080
ggagacagag gttgcagtga gccaagatca caccactgca ctccagcctg gttgacagag  43140
cgagactctg tctcaaaaaa aaaaaaatcc catttgctca tttttggat actagtataa  43200
ctatcactct aaaccagtta gtacttaaat caagcagata tgggagatgg tgaattacca  43260
tctacagtgt tgtcatatat gtcacatact gagcattatc agctagtaga atctagttaa  43320
ttgttctatg tgtgatgtat gcagagttcc cattttgaat gtgtttttac tatgcttaaa  43380
```

-continued

```
taaatgactg atgtcagcaa ccccaaaatg atacatctga tgtaagagcc cctgttcccc  43440
aataataaca tctaaactat agacattgga atgaacaggt gcccctaagt ttcctccctc  43500
cagggtttct tggccggtct ctgaggacta cacatcccta ctcccgtctt tcctcatctt  43560
caggcgcagt aacagtatct ccaagtcccc tggccccagc tccccaaagg agcccctgct  43620
gttcagccgt gacatcagcc gctcagaatc ccttcgttgt tccagcagct attcacagca  43680
gatcttccgg ccctgtgacc taatccatgg ggaggtcctg ggaagggct tctttgggca   43740
ggctatcaag gtgagcgcag gcaacaattg ctttgctctt ctgcccccag tccctctgtc  43800
actgtctttc ggggatttct catcacttgg ccccacccca caccatgcag gatgccaggc  43860
ctccttcctg gctttgggtg ttggtgtgag aggtatcctt caccccccacc caggccacct 43920
aaggtcaatg ttgctgttac agtgagcttg tggacctgga gatccaggtt gggttgagct  43980
gtgcctgtgg ccctcctgcc tccagtcagt gggtgtttgt taggtgcctg cagacctcag  44040
taccgggcat gctacaagga gcacacaggg gaatggctcc tgcctccctg gtgaacagtc  44100
tcagggacta acctctctct ttctctcctc ctcctcctct tctgctgaga actgggaggg  44160
ggggtcaggt aagacgtgtg tctcagcttg ggggcagcag ggctggagag ctcacccccg  44220
atccacccag ctccctggtg catgtctttg gcactgacct tcctgccccc agacttctgt  44280
tcactcagga gactcacttc tatgccaaat gaccagagcc cctgcttggc ttggcagcat  44340
cccctcctgc cttcttcccc acttcccttt tctgggttct tgcctgtcct ctgtgcatgc  44400
ccagctctcc aggaaagagg gtttgcttcc gtgtgagtcc catgttgctc cacgctgcat  44460
cttccacaca tgaactctgt cattctgacc cggctcagtg tgccctccaa gggatgggat  44520
ggccagctgc atagattttc tcaaacagtt ctccagaact tcctctggtc tcagcaccat  44580
taacagtcac cctccctgta ggtgacacac aaagccacgg gcaaagtgat ggtcatgaaa  44640
gagttaattc gatgtgatga ggagacccag aaaacttttc tgactgaggt aagaagatgg  44700
agggggcccg ggaggttggt gtcaccattg gaagagagaa gaccttacaa ataatggctt  44760
caagagaaaa tacagtttgg aattactgtc ttaaagacta agcagaaaag agccctagag  44820
gaatatccca ctccctctaa attacagcgt aattatttgt tcaatgaaca cttactaaaa  44880
gcaacacaaa cagggtacaa gggatgcagt aacaaaagat acagggttca gaagagctct  44940
caggttatga ggatgatgga catgaaaaca ctccaattta gtacaactca atgttataat  45000
cctcacctga acgccctgct aagggagcct ggaggggagc tccctgagca ctcacactcc  45060
ttgggcattt acagttttca ctacccctcc caagttactt catggagtaa cttaagttgg  45120
ggacacctgt ggtctgggta ttgccctcca agccacttgg ccactcccac cccagttctc  45180
ccaatgcagt tccaagggta aggcctatga agccatctcc atctatatgg tggtggtctt  45240
ccctcatcct gatcttagtg ccctgtcata tcacaagata ggaggtagga gatacaggtg  45300
gtaacacttg tcaagctgat tccttggagg gaagaggtaa ggaagacagt gagaagttaa  45360
ccaccagctt tccttggctt cccccacccc caggtgaaag tgatgcgcag cctggaccac  45420
cccaatgtgc tcaagttcat tggtgtgctg tacaaggata agaagctgaa cctgctgaca  45480
gagtacattg agggggcac actgaaggac tttctgcgca gtatggtgag cacaccaccc   45540
catagtctcc aggagccttg gtgggttgtc agacacctat gctatcacta ccctaggagc  45600
ttaaagggca gaggggccct gctttgcctc caaaggacca tgctgggtgg gactgagcat  45660
acatagggag gcttcactgg gagaccacat tgacccatgg ggcctggacc acgagtggga  45720
cagggctcaa cagcctctga aaatcattcc ccattctgca ggatccgttc ccctggcagc  45780
```

-continued

```
agaaggtcag gtttgccaaa ggaatcgcct ccggaatggt gagtcccacc aacaaacctg  45840
ccagcagggc gagagtaggg agaggtgtga gaattgtggg cttcactgga aggtagagac  45900
cccttcctat gcaacttgtg tgggctgggt cagcagctat tcattgagtt tgtctgtgtc  45960
actgaaactg accccagcca actgttctca gttcacagcc ctgttttcaa agaattacac  46020
atctctaaag gcaaacaggg cacggacaag gcaaactgga gaggcaaact gtagcctgag  46080
atggcctggg cttgccatca caggtattca ggtgctgagg gcccttagac caactagagc  46140
acctcactgc ctaggaaatc aatgaagggg aaatgagttc tagcggagcc ctgaaggatc  46200
agaattggat aaagttctta ttggcagaga ggcaccagga ttgaagtgac aggagcaaag  46260
acctgggagg aaagaggaga aaatcatcta tttcacctgg aaacaaatga ttccaagcat  46320
agaaataata acagctgaca agtactgagt gccctctata tgctaggcac tgggctgagg  46380
gattaacatg catgtgcatg tttattcctc atgacaacct tggtttccag ataagctgga  46440
ctggaaaggg acagagctgg gatcctgggc taatcagtct ggtcgccaag cctgagactt  46500
tagccactgc ccttcacatg ggggtccatg aaaatagtag tagtctggaa cagtttgggg  46560
gtacatcaag gtcgctgtgt tttaagctat ggagtctgga ctataggaga caaatgtaaa  46620
agagtttttt ggttgactgg ctttttggtt tttttgtttg tttgtttgtt tgtttgtttg  46680
tttgtttgtt ttttcctgtt tctggggctt gaatcaggaa ggaggttttt ttgttgttgt  46740
tgttttgaga aaggatattg ctctgttgcc cagactggag tgcagtggca cgatcatggc  46800
tcactacagc ttcgacctcc tgggctcaag caatcctcct gccttagcct cccaagtagc  46860
tggactacag gtgtgtacca ccacacctaa tttttttgaat tttttttttct tttttttttt  46920
tttttttttt ggtagagaca ggttctcact ttgttgccca ggcctgaatc tcaaactcct  46980
gggctcaagc attcctcctg cctcgccctc ccaaagtgtt gggattacag ttgtgagcca  47040
ccatgcccgg caggaaaaga tttttaagca agaaagctta agagctgtgg tttttccaaa  47100
atgagtctgg gctggcacag tggctcatgc ctgtaatccc agcacttttt tgggaggccg  47160
aggtgagtgg atcacttgag gtcaggagtt tgagaccagc ctggccaact ggtgaaaccc  47220
ctgtttctac taaagaaaaa aatgcaaaaa ttagctgggc gtggtggtgc acgcctgtag  47280
tcccagctac tcaggaggcc gaggcaggag aatagcttga acctgggagg cagaagttgc  47340
agtgagccaa gatcacacca ctgcattcca gcctgggtga cagagtgaga cttcatctca  47400
aaaaaaaaaa aaaagagaga ctgatatggt tagtacattg ggtggaatg cggagggtcc  47460
agggaatgga gccctgcata gggggctaat gaaacatttc agatttctga attaaggtag  47520
tggctgtggg gacaggagcc tgggaggcag ggtggagtca gaatggagag actggttggc  47580
aatgagggaa caggaggagg aggaggagga gttacgagtg gcttgaggtg tcacttacca  47640
gacatttggg ggatgggggga tagccgtgat tgttgagcaa ctggtttggg aagagctagc  47700
attgatccct gctgttctgt gctagcagaa cctatcagca tcttctgggc aggaaactgg  47760
ctccatgaga ctggcttagg gagaggctgc tagtcaccta atctgcagag aaggggcagc  47820
tggagctgtg ggacagaaga ggcatccatg tagctggtgg gggtgtctca gcttgtgaag  47880
aggagatggc tttgagcagg gctgacactg aaaaggctgg aagaaaaaaa cagacacaca  47940
agagtctcag gatcaggtag cataggaaag ttgtggacag tctttgagga gcactccctc  48000
aggcaggcag gcaggcaggt catgagctat agcgattcag gaagagctcc ctgggtgtgt  48060
gagcagctcc aggagcctaa gggatgaaag tagtattgca gggggctgga gagcaaggag  48120
```

-continued

```
tggctccttc tacatttgca agggaaggag aaaggaagtt gctcctgaga gtggtaagag  48180
tcagtggtgg aggcctggag aggagacata acaaacaaat ttgttgacaa acattttggt  48240
aggaaggggg agagcttaaa gtttagacag tggggaaggt ggagtcttag aggaggtgaa  48300
tgtctgaaag acagagctag ctggagcaag aagtcacttc tctgttgcag gcaggaagga  48360
tccaaagtgg ctcaagccag agattgggag agtggggagg agggagcagc ctggatctaa  48420
gtaaaatggg tagaggtgga gggggtgctg caacggccag ggttttctga agttggggac  48480
attaggagag agctgtgagg gctttggcca gccactgtgc tagtgattgg tgaaccaaag  48540
gatgggcagg agatggcagc agggaagcag aggaagtcca ggcttcctgt tggtattggg  48600
acaagggaga ggccatagga ggccctggcc ctgttgtcca ggttgggttc tgaagctggg  48660
tgggcatggc ctggtaggag agcatctatg gcgcccaatt ccagattcag ggtctagttg  48720
atttgctggc cctgtagcct cagctcatgc ttctgttcca ggcctatttg cactctatgt  48780
gcatcatcca ccgggatctg aactcgcaca actgcctcat caagttggta tgtcccactg  48840
ctctgggcct ggcctccagg gtcctatcct tcctggcttc cttgtcacaa aggaggctga  48900
cttgtcccct ctggctagag ggcagaggtg ttgcctagga gctcctatct ttcccttcct  48960
gcttcttcca atgcccttct ctgtcctctg ggagctccga gacacacaca gacataattt  49020
caccttctct cattagcaac ctttgaaata atttgattag aagggacttc agaagtttgt  49080
tgactatatg tagaaaaccc tgtcatttta cctgcttttg ccccatagta gtcttgtaaa  49140
acagttcatt gctgacccca ttttacagtg gtggcacctg aagcctcagc ctgaggccac  49200
cgagctagta aatttacagg gaccagtttg agaccagcat tcctcccact gcccctcagc  49260
tgtggtggtt acaatgttgt ttgtcttact gacttgctat ctggcttcct gggtgtctac  49320
cggctggccc tggctctgcc ctctagaccc acaccacgca atcttcattc ctttcccaca  49380
tgactgccct gtagctattc aaagagcttg tctcccccaa gtctccccat ctactgcctc  49440
caccttgcct ttttctgtct tatcctggtt ctagccactg cctgaaatca ttttaggaat  49500
aagacaggac agggaaaaac aaaagcaacc ccctgtccca cctctgagtt ccactctcca  49560
agtccctgag cctcacctcc agggctccag tggctctgcc atgaacccac tgtgggctgg  49620
gagtctgctg tgcacagata ccagaccctc agaaacacaa atgccaagtg tgtctgtttt  49680
tttgttttgt tttgttttgt tttttagatg gagtctcatt ctgtttccca ggctggagtg  49740
cagtggtgca atcttggctt actgcagcct ctacctcccg ggttctagtg attgttctgc  49800
ttcagcctcc cagtagctag gactacaggc gtgtgccacc acgcccagct aatttttttt  49860
tttttttttt tgtattttta gtagagacag ggttttgcca tgttggccag gctggtcttg  49920
aactcctgac ctcaggtgat tcacccgcct tggcctccca aagttctggg attacaggtg  49980
gaagccaccg tgcctggcct gagtgtgtct atttgataga gctttctgct ctgattctcc  50040
cttgctatac accttttctc cccttctcag tggcttctct tgcctatgct tcctccccag  50100
ggccaggttt gagaacatcc ccatgaagtc ctgacctgtc ttttatccta ccaggacaag  50160
actgtggtgg tggcagactt tgggctgtca cggctcatag tggaagagag gaaaagggcc  50220
cccatggaga aggccaccac caagaaacgc accttgcgca agaacgaccg caagaagcgc  50280
tacacggtgg tgggaaaccc ctactggatg gcccctgaga tgctgaacgg tgagtcctga  50340
agccctggag gggacacccg cagagggagg acagatgctg cccttgcatc agagccctgg  50400
gaattccagg ggaggcctgt gaagcgtagg accggatacc cagagctgag gatattttc  50460
ccttgccagg tggggcctca cgatttagct cctgagctca gggggctggg aactgatcag  50520
```

-continued

```
tgtcccatca tggggataa ggtgagttct gactgtggca tttgtgcctc agggatcgct  50580
aagagctcag gctattgtcc cagctttagc cttctctctc catggtgaga actgaagtgt  50640
ggtgccctct ggtggataat gctcaaacca accagagatg ctggttggga ttcttgaaat  50700
cagggttgtg aggcctcaga aatggtctga atacaatcca ttttggagtc tgaggcccag  50760
agaagttcag tgaattgcct aggagcatac agctgcctaa tggcagaggc tagatgaacc  50820
ctagtctggt tcttttccac tttaacgtgc agtttcatcc taggcagtgt tatgttataa  50880
gggctctcca aggcagttca cctacggctg aggaaggact attttcaggt ggtgtctgcg  50940
caggacagcc tgtggggtgt ccctacagaa cctgttctag ccctagttct tagctgtggc  51000
ttagattgac cctagaccca gtgcagagca ggtaagggat gtaaacttaa cagtgtgctc  51060
tcctgtgttc cccaaggaaa gagctatgat gagacggtgg atatcttctc ctttgggatc  51120
gttctctgtg aggtgagctc tggcaccaag gccatgcccg aggcagcagg cctagcagct  51180
ctgccttccc tcggaactgg ggcatctcct cctagggatg actagcttga ctaaaatcaa  51240
catgggtgta gggttttatg gtttataacg catctgcaca tctttgccac gttcgtgttt  51300
cattggtctt aagagaagga ctggcagggt tttttgtttt tagatggagc ctcacttcgt  51360
tgcccaggct ggagtgcagt ggcacaatct gggctcactg caacctctgc cttctgggtt  51420
caagtgattc tcctgcctca gcctcccaag tagctgggac taccggcaca caccaccatg  51480
cccggctaat tttgtatttt ttagtagaga cagggtttca ccatgttggc caggctggtc  51540
ttgaactccg gacctcaggt gatccgcctg cctcagcctc taaaagtgct ggaattaata  51600
ggcgtgagct acctcgcccg gccaggtttt tttttttttt tttttagttg aggaaactga  51660
ggcttggaag agggcagtgg cttgcacatg gtcgataagg ggcagatgag actcagaatt  51720
ccagaaggaa gggcaagaga ctgttcatgt ggctgtctag ctagctcttg ggccaaatgt  51780
agcccttctc agttcccttc aagtagaagt agccactcta ggaagtgtca gccctgtgcc  51840
aggtaccacg tggacagagt gaggaatctt ggaaagattc ctacctttag gagtttagtc  51900
aggtgacagc atatctcagc gactcaaaca cacacacatt caaagccttc tgtaattcct  51960
acaaagttgt gaggggtaga ggagaggaga gacaagggat ggttaggata atgaaggaat  52020
gttttgtttt tgtttttgtt tttgagatgg agtttcactc tgtcacccag gctggagtgc  52080
agaggtgcaa tcttggctca ctgcagcctc cgcctcccag gttcaagcaa tcctcctgcc  52140
tcagcctccc aagtagctgg gactacaggt gtgcgccacc acgcctggct aatttttgta  52200
ttttcagtag agacagggtt tcgccatatt ggccaggctg gtctcaaatg cctgacctca  52260
ggtgatacac ccgcttcagc ctcccaaagt gctgagatta caggcatgag ctaccgtgcc  52320
tggccatgaa ggaagatttg ttttaaaaaa ttgttttctt taatattaat tgaacacctc  52380
tgttcagagc actgggctgg tgccagaggg tttcagacat gaatcagatc cagcacctca  52440
tagagcctta atctggcaca cacacacagc cacaaggaga cacagacaag gcagggtagg  52500
atgagtggaa gctaggagca gatgctgatt tggaacactt ggcttctgca gtgaagcccc  52560
ttcttagtcc tcttcagtaa cccagctctc agtggataca ggtctggatt agtaagattt  52620
ggagagatga ttggggattg gggagagctc tctaacctat tttaccacct cctcttctgc  52680
cattcttcct gtccacatcc ccagcatccc tttcccttgc caagtatctg tggcctctgt  52740
agtcctttgt aaacagctgt cttcttaccc tacagatcat tgggcaggtg tatgcagatc  52800
ctgactgcct tccccgaaca ctggactttg gcctcaacgt gaagcttttc tgggagaagt  52860
```

-continued

```
ttgttcccac agattgtccc ccggccttct tcccgctggc cgccatctgc tgcagactgg  52920
agcctgagag caggttggta tcctgccttt ttctcccagc tcacagggtc ctgggacgtt  52980
tgcctctgtc taaggccacc cctgagccct ctgcaagcac aggggtgaga gaagccttga  53040
ggtcaagaat gtggctgtca acccctgagc catctgacaa cacatatgta caggttggag  53100
aagagagagg taaagacata gcagcaagta atctggatag gacacagaaa cacagccatt  53160
aaaagaaagt ttaaaagaag gaaattcacc caaaccattt gaatacagta agtgtattca  53220
tctttcgata ttcccctgtc catatctaca catatacttt tttttatagt aaatagttct  53280
gtattttgcc ctgcatttcc cttgtgttta ctatccagtc ttcctgttta tcatttttgt  53340
cgacaacatg aaattctatt gagagactgt ctgaacatat tgtaatgtag atgttcaggt  53400
ttttccagtt tctctttaca ataggtattt aactacagtg agcagtttta tgcatttagc  53460
taatttctcc tttgaggaag tattttcaaa attaccttta ttcttctcag gtaataattt  53520
cattattacc aaagttaccc taggtctttt caagtgtgtg gttaaaaaac gagaatctgg  53580
ctgggcgcga tggctcacac ctgtaatccc agcactttgg gaggctgagg ctggtggatc  53640
acctgaggtc tggagttcga gaccagcctg gccaacatgg tgaaacccca tctctactaa  53700
aaatacaaaa cttagccagg catggtggca ggtgcctgta accccagcta cttgggaggc  53760
tgaggcagga gaattgcttg aacccagggg cggaggttgc agtgagccga tatcacgcca  53820
ttgcactcca gcctcggcaa caagagtgaa actctgtctc aaaaatgggg ttcttttcct  53880
gccatcaaaa atcatgtttc ttttaaaaac aagttcaaac attaccaaag tttatagcac  53940
aggaaatacg tcttctgtaa tctcccttaa ccaatatatc cctcaacatt ctcctcaccc  54000
ccaactccac cctcccagga taaccagttg ggacataatc tttatttaaa aatggtttcc  54060
ggatagagaa agcgcttcgg cggcggcagc cccggcggcg gccgcagggg acaaagggcg  54120
ggcggatcgg cggggagggg gcggggcgcg accaggccag gcccgggggc tccgcatgct  54180
gcagctgcct ctcgggcgcc cccgccgccg ccctcgccgc ggagccggcg agctaacctg  54240
agccagccgg cgggcgtcac ggaggcggcg gcacaaggag gggccccacg cgcgcacgtg  54300
gccccggagg ccgccgtggc ggacagcggc accgcggggg gcgcggcgtt ggcggccccg  54360
gccccggccc ccaggccagg cagtggcggc caaggaccac gcatctactt tcagagcccc  54420
ccccggggcc gcaggagagg gcccgggctg ggcggatgat gagggcccag tgaggcgcca  54480
agggaaggtc accatcaagt atgaccccaa ggagctacgg aagcacctca acctagagga  54540
gtggatcctg gagcagctca cgcgcctcta cgactgccag gaagaggaga tctcagaact  54600
agagattgac gtggatgagc tcctggacat ggagagtgac gatgcctggg cttccagggt  54660
caaggagctg ctggttgact gttacaaacc cacagaggcc ttcatctctg gcctgctgga  54720
caagatccgg gccatgcaga agctgagcac accccagaag aagtgagggt ccccgaccca  54780
ggcgaacggt ggctcccata ggacaatcgc taccccccga cctcgtagca acagcaatac  54840
cgggggaccc tgcggccagg cctggttcca tgagcagggc tcctcgtgcc cctggcccag  54900
gggtctcttc ccctgccccc tcagttttcc acttttggat ttttttattg ttattaaact  54960
gatgggactt tgtgttttta tattgactct gcggcacggg ccctttaata aagcgaggta  55020
gggtacgcct ttggtgcagc tcaaaaaaaa aaaaaaaaat gatttccagc ggtccacatt  55080
agagttgaaa ttttctggtg ggagaatcta taccttgttc ctttataggc caaggaccgc  55140
agtccttcag taacaccagt gtaaaagctt gaggagaaat tgtgaagcta cacagtattt  55200
gttttctaat acctcttgtc attctaaata tctttaattt attaaaaaat atatatatac  55260
```

-continued

```
agtattgaat gcctactgtg tgctaggtac agttctaaac acttgggtta cagcagcgaa  55320
caaaataaag gtgcttaccc tcatagaaca tagattctag catggtatct actgtatcat  55380
acagtagata caataagtaa actatattga atattagaat gtggcagatg ctatggaaaa  55440
agagtcaaga caagtaaaga cgattgttca gggtaccagt tgcaatttta aatatggtcg  55500
tcagagcagg cctcactgag gtgacatgac atttaagcat aaacatggag gaggaggagt  55560
aagcctgagc tgtcttaggc ttccggggca gccaagccat ttccgtggca ctaggagcct  55620
ggtgtttccg attccacctt tgataactgc attttctcta agatatggga gggaagtttt  55680
tctcctattg tttttaagta ttaactccag ctagtccagc cttgttatag tgttacctaa  55740
tctttatagc aaatatatga ggtaccggta acattatgcc catttctcac agaggcacta  55800
ctaggtgaag gagtttgcct gacgttatac aaccaggaag tagctgagcc tagatccctt  55860
ccacccaccc catggccctg ctcatgttcc acctgcctct aatttacctc ttttccttct  55920
agaccagcat tctcgaaatt ggaggactcc tttgaggccc tctccctgta cctgggggag  55980
ctgggcatcc cgctgcctgc agagctggag gagttggacc acactgtgag catgcagtac  56040
ggcctgaccc gggactcacc tccctagccc tggcccagcc ccctgcaggg gggtgttcta  56100
cagccagcat tgcccctctg tgccccattc ctgctgtgag cagggccgtc cgggcttcct  56160
gtggattggc ggaatgttta gaagcagaac aagccattcc tattacctcc ccaggaggca  56220
agtgggcgca gcaccaggga aatgtatctc cacaggttct ggggcctagt tactgtctgt  56280
aaatccaata cttgcctgaa agctgtgaag aagaaaaaaa cccctggcct ttgggccagg  56340
aggaatctgt tactcgaatc cacccaggaa ctccctggca gtggattgtg ggaggctctt  56400
gcttacacta atcagcgtga cctggacctg ctgggcagga tcccagggtg aacctgcctg  56460
tgaactctga agtcactagt ccagctgggt gcaggaggac ttcaagtgtg tggacgaaag  56520
aaagactgat ggctcaaagg gtgtgaaaaa gtcagtgatg ctcccccttt ctactccaga  56580
tcctgtcctt cctggagcaa ggttgaggga gtaggttttg aagagtccct taatatgtgg  56640
tggaacaggc caggagttag agaaagggct ggcttctgtt tacctgctca ctggctctag  56700
ccagcccagg gaccacatca atgtgagagg aagcctccac ctcatgtttt caaacttaat  56760
actggagact ggctgagaac ttacggacaa catcctttct gtctgaaaca aacagtcaca  56820
agcacaggaa gaggctgggg gactagaaag aggccctgcc ctctagaaag ctcagatctt  56880
ggcttctgtt actcatactc gggtgggctc cttagtcaga tgcctaaaac attttgccta  56940
aagctcgatg ggttctggag gacagtgtgg cttgtcacag gcctagagtc tgagggaggg  57000
gagtgggagt ctcagcaatc tcttggtctt ggcttcatgg caaccactgc tcacccttca  57060
acatgcctgg tttaggcagc agcttgggct gggaagaggt ggtggcagag tctcaaagct  57120
gagatgctga gagagatagc tccctgagct gggccatctg acttctacct cccatgtttg  57180
ctctcccaac tcattagctc ctgggcagca tcctcctgag ccacatgtgc aggtactgga  57240
aaacctccat cttggctccc agagctctag gaactcttca tcacaactag atttgcctct  57300
tctaagtgtc tatgagcttg caccatattt aataaattgg gaatgggttt ggggtattaa  57360
tgcaatgtgt ggtggttgta ttggagcagg gggaattgat aaaggagagt ggttgctgtt  57420
aatattatct tatctattgg gtggtatgtg aaatattgta catagacctg atgagttgtg  57480
ggaccagatg tcatctctgg tcagagttta cttgctatat agactgtact tatgtgtgaa  57540
gtttgcaagc ttgctttagg gctgagccct ggactcccag cagcagcaca gttcagcatt  57600
```

-continued

```
gtgtggctgg ttgtttcctg gctgtcccca gcaagtgtag gagtggtggg cctgaactgg  57660

gccattgatc agactaaata aattaagcag ttaacataac tggcaatatg gagagtgaaa  57720

acatgattgg ctcagggaca taaatgtaga gggtctgcta gccaccttct ggcctagccc  57780

acacaaactc cccatagcag agagttttca tgcacccaag tctaaaaccc tcaagcagac  57840

acccatctgc tctagagaat atgtacatcc cacctgaggc agccccttcc ttgcagcagg  57900

tgtgactgac tatgaccttt tcctggcctg gctctcacat gccagctgag tcattcctta  57960

ggagccctac cctttcatcc tctctatatg aatacttcca tagcctgggt atcctggctt  58020

gctttcctca gtgctgggtg ccacctttgc aatgggaaga aatgaatgca agtcacccca  58080

ccccttgtgt ttccttacaa gtgcttgaga ggagaagacc agtttcttct tgcttctgca  58140

tgtgggggat gtcgtagaag agtgaccatt gggaaggaca atgctatctg gttagtgggg  58200

ccttgggcac aatataaatc tgtaaaccca aaggtgtttt ctcccaggca ctctcaaagc  58260

ttgaagaatc caacttaagg acagaatatg gttcccgaaa aaaactgatg atctggagta  58320

cgcattgctg gcagaaccac agagcaatgg ctgggcatgg gcagaggtca tctgggtgtt  58380

cctgaggctg ataacctgtg gctgaaatcc cttgctaaaa gtccaggaga cactcctgtt  58440

ggtatctttt cttctggagt catagtagtc accttgcagg gaacttcctc agcccagggc  58500

tgctgcaggc agcccagtga cccttcctcc tctgcagtta ttcccccttt ggctgctgca  58560

gcaccacccc cgtcacccac cacccaaccc ctgccgcact ccagccttta acaagggctg  58620

tctagatatt cattttaact acctccacct tggaaacaat tgctgaaggg gagaggattt  58680

gcaatgacca accaccttgt tgggacgcct gcacacctgt ctttcctgct tcaacctgaa  58740

agattcctga tgatgataat ctggacacag aagccgggca cggtggctct agcctgtaat  58800

ctcagcactt tgggaggcct cagcaggtgg atcacctgag atcaagagtt tgagaacagc  58860

ctgaccaaca tggtgaaacc ccgtctctac taaaaataca aaaattagcc aggtgtggtg  58920

gcacatacct gtaatcccag ctactctgga ggctgaggca ggagaatcgc ttgaacccac  58980

aaggcagagg ttgcagtgag gcgagatcat gccattgcac tccagcctgt gcaacaagag  59040

ccaaactcca tctcaaaaaa aaaaa                                         59065
```

```
<210> SEQ ID NO 4
<211> LENGTH: 265
<212> TYPE: PRT
<213> ORGANISM: Human

<400> SEQUENCE: 4

Leu Thr Glu Val Lys Val Met Arg Ser Leu Asp His Pro Asn Val Leu
1               5                   10                  15

Lys Phe Ile Gly Val Leu Tyr Lys Asp Lys Lys Leu Asn Leu Leu Thr
            20                  25                  30

Glu Tyr Ile Glu Gly Gly Thr Leu Lys Asp Phe Leu Arg Ser Met Asp
        35                  40                  45

Pro Phe Pro Trp Gln Gln Lys Val Arg Phe Ala Lys Gly Ile Ala Ser
    50                  55                  60

Gly Met Ala Tyr Leu His Ser Met Cys Ile Ile His Arg Asp Leu Asn
65                  70                  75                  80

Ser His Asn Cys Leu Ile Lys Leu Asp Lys Thr Val Val Val Ala Asp
                85                  90                  95

Phe Gly Leu Ser Arg Leu Ile Val Glu Glu Arg Lys Arg Ala Pro Met
            100                 105                 110
```

-continued

```
Glu Lys Ala Thr Thr Lys Lys Arg Thr Leu Arg Lys Asn Asp Arg Lys
        115                 120                 125

Lys Arg Tyr Thr Val Val Gly Asn Pro Tyr Trp Met Ala Pro Glu Met
    130                 135                 140

Leu Asn Gly Lys Ser Tyr Asp Glu Thr Val Asp Ile Phe Ser Phe Gly
145                 150                 155                 160

Ile Val Leu Cys Glu Ile Ile Gly Gln Val Tyr Ala Asp Pro Asp Cys
                165                 170                 175

Leu Pro Arg Thr Leu Asp Phe Gly Leu Asn Val Lys Leu Phe Trp Glu
            180                 185                 190

Lys Phe Val Pro Thr Asp Cys Pro Pro Ala Phe Phe Pro Leu Ala Ala
        195                 200                 205

Ile Cys Cys Arg Leu Glu Pro Glu Ser Arg Pro Ala Phe Ser Lys Leu
    210                 215                 220

Glu Asp Ser Phe Glu Ala Leu Ser Leu Tyr Leu Gly Glu Leu Gly Ile
225                 230                 235                 240

Pro Leu Pro Ala Glu Leu Glu Glu Leu Asp His Thr Val Ser Met Gln
                245                 250                 255

Tyr Gly Leu Thr Arg Asp Ser Pro Pro
            260                 265
```

That which is claimed is:

1. An isolated nucleic acid molecule consisting of a nucleotide sequence selected from the group consisting of:

(a) a nucleotide sequence that encodes an amino acid sequence shown in SEQ ID NO:2;

(b) a nucleic acid molecule consisting of the nucleic acid sequence of SEQ ID NO:1;

(c) a nucleic acid molecule consisting of the nucleic acid sequence of SEQ ID NO:3; and

(d) a nucleotide sequence that is completely complementary to a nucleotide sequence of (a)–(c).

2. A nucleic acid vector comprising a nucleic acid molecule of claim 1.

3. A host cell containing the vector of claim 2.

4. A process for producing a polypeptide comprising culturing the host cell of claim 3 under conditions sufficient for the production of said polypeptide, and recovering the peptide from the host cell culture.

5. An isolated polynucleotide consisting of a nucleotide sequence set forth in SEQ ID NO:1.

6. An isolated polynucleotide consisting of a nucleotide sequence set forth in SEQ ID NO:3.

7. A vector according to claim 2, wherein said vector is selected from the group consisting of a plasmid, virus, and bacteriophage.

8. A vector according to claim 2, wherein said isolated nucleic acid molecule is inserted into said vector in proper orientation and correct reading frame such that the protein of SEQ ID NO:2 may be expressed by a cell transformed with said vector.

9. A vector according to claim 8, wherein said isolated nucleic acid molecule is operatively linked to a promoter sequence.

* * * * *